

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

A Thesis

Presented to the Faculty

of the Department of Mathematics and Computer Science

McAnulty College and Graduate School of Liberal Arts

Duquesne University

in partial fulfillment of

the requirements for the degree of

Masters of Science in Computational Mathematics

by

Sara M. Bennett

April 16, 2004

Sara M. Bennett

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

Master of Science in Computational Mathematics

Department of Mathematics & Computer Science
Duquesne University, Pittsburgh, PA, USA

April 16, 2004

APPROVED

John Kern, Ph.D., Assistant Professor
Department of Mathematics & Computer Science

APPROVED

Richard Pro, Vice President
Highmark Healthcare Informatics

APPROVED

Frank D'Amico, Ph.D., Chair
Department of Mathematics & Computer Science

APPROVED

Kathleen Taylor, Ph.D., Graduate Director of Computational Mathematics
Department of Mathematics & Computer Science

APPROVED

Constance D. Ramirez, Ph.D., Dean
McAnulty College and Graduate School of Liberal Arts

Acknowledgements

First, I would like to thank my adviser, Dr. John Kern, for all his time, patience and guidance throughout the entire process. With his support, I was able to produce a work of which I can be proud. I would also like to thank all of my committee members, especially Richard Pro for his endless support and encouragement.

Special thanks goes to Tom Schultz who gave me the inspiration for this project. His extensive knowledge of the health insurance industry is invaluable.

I would also like to thank my fellow graduate students, Melissa, Jen and Joe, for their support. I'm not sure I would have survived this process without them.

Finally, I would like to thank my family—my parents, Beth, and Joe—for always believing in me.

Contents

- 1 Introduction** **1**
- 1.1 Statement of Problem 1
- 1.2 Methodology 4
 - 1.2.1 Predictor Variables 6
 - 1.2.2 Data Collection and Cleansing 8
- 2 Data Model and Parameter Estimation** **10**
- 2.1 Data Model 10
- 2.2 Parameter Estimation 11
 - 2.2.1 Expert Opinion Gamma Point Mass Distribution 12
 - 2.2.2 EOGPM Parametrization 14
- 3 Applications** **15**
- 3.1 Results 15
- 4 Discussion** **21**
- 4.1 Summary of Model Relevance 21
- 4.2 Convergence 24
- 4.3 Future Analyses 25
- A Initial Values: $\beta_i^{(0)}$** **27**
- B Expert Values** **28**
- C Table of Estimates** **29**
- D Actual Proportion of Members Who Left When $X_{ij} = 1$** **31**

Chapter 1

Introduction

In this research, we implement a multiple logistic regression model in which the coefficients of indicator variables are constrained to be zero or positive. By doing this, the contribution of each variable to the failure probability can be assessed. Due to this restriction on the coefficients, a Bayesian approach to parameter estimation—which assigns mixture priors to the coefficients—is taken. The data is provided by a large health insurance company in Western Pennsylvania and includes the enrollment status and corresponding values of 84 predictor variables for 1,280,612 individuals. The insurer feels the analysis is needed to determine why its membership is declining, why its cost trend is higher than the national average, and what logical steps can be taken to reverse the current trends.

1.1 Statement of Problem

Over the last few years the health insurance market in Western Pennsylvania has experienced several changes. Foremost of these was the introduction of a new regional insurance provider. Before this plan was created, one insurer provided coverage for a large majority of the commercial health insurance market. Since the new plan was introduced, the large insurer has lost significant membership. In addition to the in-

creased competition, the economic impact of September 11th has caused changes in the commercial health insurance market. In an effort to control health insurance premium increases, employers, especially small business owners, have shifted more of the cost of health insurance to their employees. This cost shifting has come in the form of premium shifting as well as benefit cost shifting. For the employee, this means larger amounts deducted from their earnings to pay for insurance (the premium shifting), and higher co-pays and deductibles to pay when they need healthcare (the benefit cost shifting) (Draper et al. 2003). Along with the changing market environment, the demographics in Western Pennsylvania continue to change. Based on the average age of its residents, Allegheny County is the second oldest county in the United States. Between 1990 and 2000, the Pittsburgh metropolitan region's population decreased by 2 percent (U.S. Bureau of the Census). The age segments with the largest decreases were 60-64, 25-34, and 20-24 (-25%, -24%, and -18%, respectively). The age segments with the largest increases were 85+, 45-54, and 75-84 (41%, 38%, and 23%, respectively).

As a result of the changing market, consumers are offered more options when choosing their health insurance. In an effort to more closely compete with the newer health insurance provider, the large insurer has made more health insurance product options available to the consumer. This increase in choice has allowed the consumer to choose the health plan that best fits his or her needs and financial situation. As a result, the healthiest members generally choose the products with the lowest premiums while the members with higher disease risk choose more generous benefit plans with higher premiums but lower co-pays and deductibles that they will be required to pay throughout the year (Sutton et al. 2002). This situation leads to adverse selection, which occurs when "the people who sign up for an insurance plan have costs that are greater than the expected costs that the insurance plan used to calculate the premium" (HealthInsurance.info 2002). Adverse selection increases the

uncertainty in the pricing of products as well as causing unforeseen large and sharp changes in the underlying prices and pricing trends. This in turn, forces the insurer to raise premiums on the whole. The new increase in price (both premium and benefits) will force additional selection as the relatively lower cost people look for relief from premium increases. As the total number of members decreases, the average cost of the plan then increases, as the total health care expense (both fixed and variable costs) is spread over fewer members. This cycle of

- Relatively low cost members disenrolling,
- Average cost per member increasing, and then
- Premiums increasing

is sometimes called the premium death spiral. Adverse selection happens more often in health insurance than other types of insurance such as auto or homeowners because unlike these other types of insurance, health insurance is not required or mandated by the government or other large institutions. Since everyone, not just the drivers who have accidents, are required to have auto insurance, the average risk of an auto insurer remains fairly constant over time.

These changes have caused the large insurance provider to experience changes in their membership base. The insurer has experienced large decreases in enrollment and a surge in costs that the company had never seen before. As a non-profit organization, the insurer is required to insure anyone who requests coverage. Historically, the insurer's market share was large enough that the ratio of high risk to low risk members was stable and the cost of high risk members did not significantly influence the average cost per member. As the enrollment declines, the ratio of low risk to high risk members decreases causing the overall average cost per member to increase faster than expected. Even if the large insurer's population mix of high risk and low risk

members had not changed, the decrease in enrollment would force the insurer to raise premiums for the remaining members to cover the fixed operating costs.

If the insurer could predict with some accuracy which individuals will leave and/or join the plan, it could develop marketing strategies aimed at the people likely to leave as well as accurately set premiums for the remaining people. If the premiums are calculated with more accuracy, the degree of adverse selection will decrease and the death spiral may stop. If, however, the insurer does not develop a method for controlling its turnover, it will continue down the premium death spiral until the only members left are the ones with the highest costs. We will attempt to model the adverse selection for the large insurer by determining which demographic and health factors are most common in the members who have left.

1.2 Methodology

Member turnover is defined as members enrolling in and disenrolling from the provider's health insurance plans. To determine the characteristics of members who disenroll, we will examine a data set containing various demographic and health related variables for members who were enrolled as of June 2002. Each member in the database will be defined as a success if they are still enrolled as of June 2003. Those that left this provider before June 2003 will be defined as failures.

Let Y_i represent the success/failure status of the i th member. Thus $Y_i = 1$ if the i th member is not enrolled as of June 2003, otherwise $Y_i = 0$. Assuming the Y_i 's are independent, Bernoulli random variables with individual failure probabilities p_i , the joint distribution f of all Y_i 's is given by the product of n Bernoulli mass functions:

$$f(Y_1, \dots, Y_n | p_1, \dots, p_n) = \prod_{i=1}^n p_i^{Y_i} (1 - p_i)^{1 - Y_i} . \quad (1.1)$$

To determine how the p_i 's are linked to the Y_i 's, we assume each Y_i is accompanied

by a vector \mathbf{X}_i of covariate values

$$\mathbf{X}_i = \{X_{i1}, X_{i2}, \dots, X_{im}\},$$

where X_{ij} represents for person i the value of the j th covariate ($j \leq m$). Recognizing that each person has both a response Y and a vector of covariate values \mathbf{X} , we let the probability of a failure response ($Y = 1$) be a function of the covariate vector. Specifically, if Y_i is the response for person i , and \mathbf{X}_i is the corresponding covariate vector for this person, then we define the probability p_i that $Y_i = 1$ as

$$p_i = \frac{e^{g(\mathbf{X}_i, \boldsymbol{\beta})}}{1 + e^{g(\mathbf{X}_i, \boldsymbol{\beta})}}, \quad (1.2)$$

where

$$g(\mathbf{X}_i, \boldsymbol{\beta}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_m X_{im}. \quad (1.3)$$

In this way, we have constructed a linear model for estimating the natural logarithm of the quantity $\frac{p_i}{1-p_i}$. This natural logarithm, as a function of p_i is referred to as the *logit* function, or $\text{logit}(p_i)$. Notice that the only unknown on the r.h.s of (1.2) is the vector of parameters $\boldsymbol{\beta}$. We estimate these parameters (and hence the probability p_i for a given covariate vector \mathbf{X}_i) based on the linear regression model

$$\text{logit}(p_i) = g(\mathbf{X}_i, \boldsymbol{\beta}).$$

Before providing the details behind estimating the parameter vector $\boldsymbol{\beta}$, we first define the variables that compose the covariate vector, \mathbf{X}_i . Because the values of these variables will potentially predict the failure probability, we will also refer to these variables as predictor variables.

1.2.1 Predictor Variables

By estimating the probability of disenrollment for a given set of predictor variables, the insurer will be able to identify which members will leave the plan each year. By having a better projection of the total membership, the insurer will be able to set more accurate premiums, reduce the degree of adverse selection, and possibly stop the death spiral. We will attempt to determine which of the variables listed below are most influential in predicting turnover. In order to capture the best subset of predictor variables, we will include a wide variety of demographic, health related, and employer related variables. We give consideration to the following and point out the variable 1 through 7 are continuous variables while variables numbers 8 or greater are indicator variables (or a set of related indicator variables):

1. **HealthRisk**¹: Indexed variables for the health status of an individual. The index for both the current time period and the predictive index are included.
2. **Copay**: The amount the individual was required to pay for services between July 2001 and June 2002.
3. **Elig**: The number of months an individual was enrolled between July 2001 and June 2002.
4. **NetPayments**: The amount the insurer paid for services on behalf of an individual between July 2001 and June 2002.
5. **Age**: The age of the individual as of January 2002.
6. **Product**: The product the individual was enrolled in as of June 2002.
7. **NumProducts**: The number of products the individual had to choose from when selecting their level of coverage.

¹**HealthRisk** and **ACC**'s are based on copyrighted software developed by DxCG, Inc., 25 Kingston Street, Suite 200, Boston, MA 02111.

8. **ACC001-ACC026:** Aggregated medical condition categories. Twenty-six indicator variables that categorize the diagnoses reported for the individual between June 2001 and July 2002. A value of 1 indicates that the individual was grouped into that condition category. An individual can be assigned to more than 1 of these categories.
9. **Child:** A value of 1 for this variable indicates that the individual is under the age of 18.
10. **ClientLeft:** A value of 1 indicates that the employer cancelled their contract with the insurer.
11. **County:** The county the individual resides in. The sample is limited to the 29 counties the insurer uses to identify its Western PA region. A value of 1 for any of the 29 indicator variables indicates that the member lives in that county.
12. **Dependent:** A value of 1 indicates that the individual is an employee's dependent.
13. **Family:** A value of 1 indicates that the individual is part of a family contract.
14. **Male:** A value of 1 indicates the individual is male.
15. **Nodiag:** A value of 1 indicates that the individual had no diagnoses between July 2001 and June 2002.
16. **Nohcc:** A value of 1 indicates that the individual was not assigned to a condition category.
17. **Novalid:** A value of 1 indicates that the individual had no valid diagnosis records.

18. **Regional:** A value of 1 indicates that the employer the individual is insured through is a local company. A 0 indicates that the employer is a national company.
19. **Risk:** A value of 1 indicates that the insurer is at risk for the expenses incurred by this individual and not the employer associated with the individual.
20. **SIC:** The Standard Industry Code for the employer the individual is insured through.
21. **Spouse:** A value of 1 indicates that the individual is an employee's spouse.

1.2.2 Data Collection and Cleansing

The initial data set includes all of the predictor variables listed above for 1,280,612 individuals. In order to exclude members who may be eligible for Medicare, the sample is limited to members under the age of 60. Any negative values in **NetPayments** were set to zero. Such negative amounts occur occasionally and erroneously in the data due to the timing of adjustments. Another option would have been to eliminate the records with negative values, decreasing the number of observations in the data set.

Descriptive statistics were examined for each of the 84 independent variables. Since less than 1% of the individuals had a value of 1 for the indicator **Novalid**, it was eliminated. Several variables were removed based on their high correlation with other variables. The **SIC** indicator variables were removed because they were correlated with other variables and with each other. The concurrent health risk score was also removed since it is correlated with the prospective health risk score and with **NetPayments**. This is not surprising since the concurrent score attempts to explain the current year's expenditures while the prospective score attempts to predict next year's resource consumption. Both scores are based on the individual's

current year's diagnosis history. **Nohcc** was eliminated because of its high correlation with **Nodiag**. This is not surprising either since an individual must have a diagnosis to be categorized into an **ACC**. After eliminating these 15 variables (one each for each of the 12 **SIC** codes), we are left with $m = 69$ independent variables. The 19 variables listed above increase to 69 variables when the 26 **ACC**'s are expanded into 26 separate indicator variables and indicator variables are created for all but one of the **Counties**. A member who lives in Allegheny County has $X_{ij} = 0$ for each j that corresponds to one of the 29 **County** variables.

Chapter 2 contains a detailed description of the data model and an outline of the algorithm used to estimate the model parameters. Chapter 3 applies this model to the data and details the results of the parameter estimation. Chapter 4 concludes with a discussion of the model (its usefulness/features/limitations), computational issues surrounding the implementation of the model, and suggestions for future research.

Chapter 2

Data Model and Parameter Estimation

2.1 Data Model

Let Y_1, \dots, Y_n represent the status of each individual in our data set, where $Y_i = 1$ indicates that member i disenrolled and $Y_i = 0$ indicates that member i was still enrolled in the insurer's plan at least 12 months later. We model the outcomes Y_1, \dots, Y_n as independent Bernoulli random variables with failure probabilities p_1, \dots, p_n . This gives from (1.1) 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 from (1.2) gives an equivalent likelihood in terms of β :

$$L(\beta) = \prod_{i=1}^n \left(\frac{e^{g(\mathbf{X}_i, \beta)}}{1 + e^{g(\mathbf{X}_i, \beta)}} \right)^{Y_i} \left(\frac{1}{1 + e^{g(\mathbf{X}_i, \beta)}} \right)^{(1 - Y_i)}. \quad (2.1)$$

When estimating β it is easier to work with the natural log of the likelihood function.

Let $l(\boldsymbol{\beta})$ denote the natural log of the likelihood function in (2.1). After simplification we have:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^n \left[Y_i g(\mathbf{X}_i, \boldsymbol{\beta}) - \ln \left(1 + e^{g(\mathbf{X}_i, \boldsymbol{\beta})} \right) \right] .$$

2.2 Parameter Estimation

To estimate $\beta_0, \beta_1, \dots, \beta_m$, we first choose initial values $\boldsymbol{\beta}^0$:

$$\boldsymbol{\beta}^0 = (\beta_0^{(0)}, \beta_1^{(0)}, \dots, \beta_m^{(0)}) . \quad (2.2)$$

Actual initial values for the $m = 69$ variables under consideration were chosen by using the PROC LOGISTIC function in the SAS STATS software package. Initial values were chosen in this manner to decrease the burn-in period and to increase the rate of convergence. The initial values are reported in Appendix A.

Next we sample from the posterior distribution of these parameters using Markov Chain Monte Carlo techniques (Gilks et al. 1996). To do this, we first propose a new β_0 , call it β_0^p from a proposal density q . This proposed value is selected from a uniform distribution with parameters $\beta_0 - k$ and $\beta_0 + k$, where k is a tuning parameter set to $= 0.01$. The proposed value β_0^p will be accepted or rejected with probability α , where:

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

Here π is the prior distribution for $\boldsymbol{\beta}$, $[\cdot]^{(p)}$ denotes an expression evaluated at the proposed value of β_0 , and $[\cdot]^{(0)}$ denotes an expression evaluated at the current value of β_0 . In addition, $\frac{q(\boldsymbol{\beta}|\boldsymbol{\beta}^p)}{q(\boldsymbol{\beta}^p|\boldsymbol{\beta})}$ is the Hastings ratio of the probability of moving to the current state from the proposed state to that of moving in the opposite direction.

If the proposed value is accepted, β_0 is replaced with β_0^p in β . If not, β remains unchanged. This updating process is repeated for each β_j . Values are sequentially proposed and accepted/rejected for individual components of β until convergence is obtained.

As for the prior distribution on the β_i 's, we model these coefficients as a priori independent. The intercept and the slope coefficients for the continuous variables of g are assigned normal zero-mean priors allowing for any inverse relationships (i.e. negative coefficient values) between the independent variable and the outcome. All other slope coefficients of g —those that correspond to the indicator variables—are assigned mixture priors. In this way the slope parameters of the dichotomous variables are constrained to be either zero or positive, allowing for a straightforward measurement of a variable's contribution to the disenrollment probability. Other examples where mixture priors are useful include (Graves et al. 2003). The mixture prior chosen for the coefficients of the dichotomous variables in this problem is the Expert Opinion Gamma Point Mass Distribution (EGPM).

2.2.1 Expert Opinion Gamma Point Mass Distribution

Let g denote the density function of a gamma random variable, and let h represent the density function of a beta random variable:

$$\begin{aligned} g(x|\alpha, \beta) &= \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad \text{for } x > 0 \\ h(x|\alpha, \beta) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1}, \quad \text{for } 0 < x < 1. \end{aligned}$$

Let G and H denote the cumulative distribution functions of the gamma and beta distributions, respectively:

$$G(x) = \int_0^x g(t|\alpha, \beta) dt \quad \text{and} \quad H(x) = \int_0^x h(t|a, b) dt .$$

Now, we define the Gamma-Point-Mass (GPM) distribution. The GPM distribution is a combination of a gamma distribution and a point mass on zero. If X has a GPM distribution, then we denote the distribution as $gpm(x)$ and define it as:

$$gpm(x|\alpha, \beta, p) = \begin{cases} p & \text{if } x = 0 \\ (1 - p) \cdot g(x|\alpha, \beta) & \text{if } x > 0 \end{cases} .$$

If $GPM(x)$ represents the CDF of this distribution, then using the definition of the GPM distribution we can write $GPM(x)$ in terms of $G(x)$:

$$\begin{aligned} GPM(x) &= gpm(x = 0|\alpha, \beta, p) + \int_0^x gpm(t|\alpha, \beta, p) \cdot I_{t>0} dt \\ &= p + (1 - p) * G(x) . \end{aligned}$$

Let $U \sim beta(a, b)$. Then Y has an Expert-Opinion GPM (EGPM) when $Y = GPM^{-1}(U)$. The distribution of Y can easily be expressed from the CDF $F_Y(y)$ of Y :

$$\begin{aligned} F_Y(y) = P\{Y \leq y\} &= P\{U \leq GPM(y)\} = H(GPM(y)) \\ &= H(p + (1 - p) \cdot G(y)) . \end{aligned}$$

From this CDF we can find the EGPM density $f_Y(y)$. First consider the case when $y > 0$:

$$\begin{aligned} f_Y(y) = F'_Y(y) &= H'(GPM(y)) \cdot GPM'(y) \\ &= h(p + (1 - p) \cdot G(y)) \cdot (1 - p) \cdot g(y) . \end{aligned}$$

The probability that $Y = 0$ is found by:

$$P\{Y = 0\} = F_Y(0) = H(GPM(0)) = H(p) .$$

Combining these two results gives the EGPM distribution of f :

$$f(y|\alpha, \beta, a, b, p) = \begin{cases} H(p) & \text{if } y = 0 \\ h(p + (1 - p) \cdot G(y)) \cdot (1 - p) \cdot g(y) & \text{if } y > 0 \end{cases}$$

2.2.2 EOGPM Parametrization

The expert opinion gamma point mass distribution has gamma parameters v and w , beta parameters a and b and a probability mass of p located at zero. In order to incorporate the expert opinion, the parameters of the beta distribution a and b are set to:

$$a = qM + 1 \quad \text{and} \quad b = (1 - q)M + 1$$

where q is specified by the expert and M , whose magnitude is proportional to the strength of the expert's specification, is either fixed or variable. The vector q represents the expert's opinion on the importance of each independent variable. The expert assigned a value between 0 and 1 where a 1 is used to identify the variables he or she feels are most important. The EOPGM parameters, v, w, p , and M were fixed:

$$v = 0.1, \quad w = 10, \quad p = 0.2, \quad \text{and} \quad M = 4.$$

The vector of q values was chosen to reflect the views of the author. The values are recorded in Appendix B.

Because the data set is so large, $n = 1,280,612$, this or any other reasonable prior distribution will have negligible influence on parameter inference.

Chapter 3

Applications

The results examined here are based on the output from 12,112 iterations of the Bayesian multiple logistic regression model described in Chapter 2. The values from the first 7,609 iterations are used for burn in, which leaves 4,503 to be used for parameter inference calculations. The convergence of the model will be discussed in Chapter 4. To achieve tolerable levels of computational efficiency—and allow for a future validation study—the data was partitioned into development and testing sets. The development set contains 70% of the data (sampled randomly) or information on 896,428 individuals. The proportion of members who left is consistent across the samples; in each data set approximately 22% of the members disenrolled.

3.1 Results

Variables with a large proportion of coefficient estimates equal to zero are least likely to contribute to the probability of an individual disenrolling. The three variables that have $\geq 75\%$ of their coefficient estimates equal to 0 are listed here:

- The coefficient for **Child** has $E(\beta_i)$ of 0.00009. The probability that this $\beta_i = 0$ is 81.2%. So, being under the age of 18 has no impact on whether the member

disenrolls.

- The coefficient for **Family** has $E(\beta_i)$ of 0.00018. The probability that this $\beta_i = 0$ is 80.6%. Belonging to a family contract does not increase your probability of disenrolling. The inverse is not necessarily true since the coefficient estimate was constrained to be greater than or equal to 0. In other words, not belonging to a family contract could increase the probability of disenrollment.
- The coefficient for **Regional** has $E(\beta_i)$ of 0.00012. The probability that this $\beta_i = 0$ is 76.4%. This result implies that working for a local employer does not increase the probability of disenrolling. Again, the inverse is not necessarily true.

Variables with no coefficient estimates equal to zero are most likely to contribute to the probability that an individual will disenroll. In addition, the magnitude of their effect on the probability of disenrolling is measured by the mean value of the β estimates. The variables with all (or nearly all) of their coefficient estimates greater than zero are given here.

- **ClientLeft** has the most influence on the probability that a member will disenroll. For this variable $E(\beta_i)$ is 1.91 and $P(\beta_i > 0) = 1$. This result seems obvious since the individual will no longer have the option of being insured through this provider. There are cases, however, where the individual continues their coverage. This could be due to a variety of reasons such as:
 - the individual changes jobs, or,
 - the employer no longer offers any health insurance benefit, forcing the employee to buy his or her own policy.
- Several **County** variables have a significant impact on the probability that a member will disenroll. For each of these counties the $P(\beta_i > 0) = 1$:

County	$E(\beta_i)$	County	$E(\beta_i)$
Mercer	1.81	Warren	0.51
Beaver	0.83	Somerset	0.48
Venango	0.66	Lawrence	0.10

Further investigation should be done to determine why these counties have a greater impact on disenrollment than others. One possible explanation is that most of the providers in these areas are affiliated with the competitor.

- There are several condition categories, **ACC**'s, that increase the probability that an individual will disenroll. For all of these conditions the $P(\beta_i > 0) = 1$:

ACC	$E(\beta_i)$	ACC	$E(\beta_i)$
Cardio-Respiratory Disease	1.16	Substance Abuse	1.01
Vascular Disease	1.15	Mental Disorders	0.14

Again, further investigation should be conducted on these variables. A few possible explanations for disenrollment in these cases are that the person became disabled (a debilitating mental disorder or heart condition) or passed away (substance abuse or cardio-respiratory disease).

- **Risk** indicates whether the employer or the insurer is at risk for the health care expenditures associated with that employer and, hence, how the premiums are calculated. The individuals for which the insurer is at risk are slightly more likely to leave, $E(\beta_i) = 0.009$ and $P(\beta_i > 0) = 0.998$. This could be attributed to the price the insurer sets for these individuals.
- An individual who has no diagnoses during a 12 month period, **Nodiag**=1, is more likely to leave. The mean coefficient estimate for this variable is 0.006 and the $P(\beta_i > 0) = 0.992$.

The coefficient estimates for the continuous variables are interpreted differently, as these coefficients were allowed to range over the whole set of real numbers (positive and negative values). Each continuous covariate has a different range of plausible values. Hence, the range of values has to be combined with the mean value of the estimate to determine the effect these variables have on the probability of disenrolling. The following is a list of the continuous variables that are most likely (based on the proportion of coefficient estimates equal to zero and the mean value of the estimates) to influence the failure probability.

- **NetPayments** has an inverse relationship with the probability that an individual will leave. The expected value of β for this variable is very small, -0.00001 , but never assumes a value of zero. The values, X_{ij} for this covariate range from \$0 to over \$1,000,000 with a mean value of \$1,450 and median of \$312. The range of values the coefficient estimate could take for this variable (as compared to those for a zero-one variable) range from 0 to less than -10. A person with mean annual expenditures would have a coefficient equivalent to -0.01. This result implies that a person with higher annual expenditures is less likely to disenroll than someone with little to no costs.
- The **HealthRisk** variable has an inverse relationship with the disenrollment probability. The expected value of β for this variable is -0.012, with the vast majority of these estimates less than zero. The prospective health risk score for this sample ranges from 0.088 (healthy) to 89.251 (very ill) with a mean value of 0.97. Because the coefficient for **HealthRisk** is negative, a person with a high risk score is more likely to disenroll.
- **NumProducts**, the number of products a member has to choose from, also has an inverse relationship with the probability of disenrolling. The number of products offered ranges from 1 to 10, with a mean of 2.96 and median of 3.

Members who had fewer products to choose from are more likely to disenroll than those with multiple offerings ($E(\beta_i) = -0.13$). Combining the range of values with the mean value of the coefficient estimate gives a possible range of -0.13 to -1.3. An individual offered the average number of products would have a coefficient equivalent to -0.39. The number of products offered can be an indication of the size of the employer, larger employers would be able to offer more selection than smaller employers (where size is measured by the number of employees).

- **Age** also has a significant impact on the probability that a person will leave, $E(\beta_i) = -0.01$. Only members under the age of 60 were included in the data. The mean age of this population is 31, the median is 33. The range of values the coefficient could take (as compared to the coefficients of the zero-one variables) are 0 to -0.60. This means that a younger person is more likely to disenroll than an older person. The equivalent coefficient for an average member is -0.31.

The relevance of these results to the problems the insurer is facing will be discussed further in Chapter 4.

A complete list of the mean estimate for each β and the corresponding proportion of coefficient estimates that were equal to 0 is reported in Appendix C. The first variable listed in Appendix C is the intercept term, the next 7 are the continuous variables, the next 8 are dichotomous variables listed in decreasing order by the number of individuals who possess that characteristic, the next 26 are the individual **ACC**'s, and the last 28 are the individual **County** variables. The **ACC** and **County** variables are also listed in descending order by prevalence.

In order to corroborate the results for the dichotomous variables, the response variable was examined for all individuals corresponding to cases where $X_{ij} = 1$ for each dichotomous variable $j = \{8, 9, \dots, 69\}$. The proportion of individuals in the data set who disenrolled when $X_{ij} = 1$ is reported in Appendix D. We would expect that the

variables with the largest proportion of members who disenrolled when $X_{ij} = 1$ would have the most impact in the model. For instance, the variable **ClientLeft** contains the largest proportion (56%) of disenrolled members when **ClientLeft** = 1. The coefficient estimates from the implementation of our model tell us that **ClientLeft** has the largest impact on the probability of leaving. Another example of the corroboration: Of the individuals classified as children (under the age of 18), 21.37% had a response variable equal to 1. The proportion of the total sample that left is 21.94%, which is close to the proportion of children who left. The coefficient estimate for **Child** confirms that **Child** has no influence on the probability of disenrolling.

Chapter 4

Discussion

4.1 Summary of Model Relevance

In this research, we have applied a multiple logistic regression model to individual enrollment data obtained from a large health insurance provider. The model allowed

- coefficients for continuous variables to range over the set of real numbers (both positive and negative), but
- restricted coefficients of dichotomous variables to be greater than or equal to zero.

Such a restriction allows for measuring the influence each covariate has on the disenrollment probability, both in terms of the probability of contributing to disenrollment and the magnitude of the contribution. This coefficient specification is easily modelled using a Bayesian approach. The choice of an EOPGM prior distribution on the zero-or-positive coefficients allows for the expert opinion to be incorporated in a straightforward manner. (Although, in this application, the data set is large enough to dominate the prior opinion and hence the inference on β . The contribution of the prior to the log-likelihood sum, as given in Equation 2.2, is $< 1\%$.)

The results of this model indicate that the insurer's commercial managed care business may be in a premium death spiral. Recall the death spiral is defined by the cycle of relatively lower cost members disenrolling, causing the average cost per member to increase, which in turn causes the insurer to raise premiums. We now ascertain evidence from our model output in support of these components:

- Several results point toward lower cost members disenrolling:
 - **NetPayments** has a significant inverse relationship with the probability of disenrollment. This implies that relatively lower cost members are more likely to leave.
 - The member's **RiskScore** has an inverse relationship with the disenrollment probability. This result implies that a healthier person (a lower risk score) is more likely to disenroll. A healthier person would have lower annual expenditures.
 - The member's **Age** also has an inverse relationship with the probability of disenrollment. Younger members typically incur lower costs than older members. Therefore, younger members having a higher probability of disenrolling implies that less expensive members are more likely to disenroll.
 - If the covariate **Nodiag** is equal to 1, the probability of disenrolling increases. Having no diagnoses implies that the member incurred very little, if any, costs during the year. Hence, this result also implies that members with low costs are more likely to disenroll.
- Low cost members disenrolling will cause the average cost per member to increase if the insurer does not enroll new low cost members to replace those that left. The insurer's base enrollment has been decreasing over the past few years so it is not probable that the new membership is offsetting the effect of lower cost members leaving.

- It is widely known that the insurer has been raising premiums by 20% or more each year. The coefficient estimate for **Risk** indicates that members the insurer is at risk for are more likely to leave. The premium amount for these individuals is based on the cost trend for all members of the insurance plan while the premium amount for members whose employer is at risk for their expenses is based solely on the expenses those employees incur. So, the effect of adverse selection would impact the premium charged to members for which the insurer is at risk, not those for which the employer is responsible.

The same model parameters would have to be estimated for previous and future time periods to be able to state definitively that the insurer is in a premium death spiral. If the results of the model are not replicated for other time periods, the adverse selection that occurred here could be considered a one time event and the insurer may not be in the midst of a death spiral. Although, other informal studies in which the author has been involved indicate that the cycle of events that define a premium death spiral started before June 2001 and promise to continue through 2004.

In order for the insurance company to use any of the results reported herein to improve profitability, many different departments would have to be convinced of the model's accuracy and dependability. As with many organizations, the majority of employees do not have an extensive statistical or mathematical background. These people may have trouble accepting results that are different from or in opposition to common beliefs about the industry.

Assuming the right people are convinced of the impact the results of this model could have, many positive actions could be taken to gain back some market share and increase profitability. One of the easier measures to implement that could increase enrollment would be for the marketing department to turn their focus to the demographic and geographic factors most likely to increase the probability of dis-enrollment. Actuaries could use the model to predict which members will still be

enrolled during the next time period so that they could set more accurate premiums. Many other marketing, medical management, and pricing policy efforts could be improved by combining the results from this model with their current processes.

4.2 Convergence

Because the data matrix examined here is so large—896,428 rows by 70 columns—the processing time for one update of β takes approximately 2.75 minutes. This time includes all of the measures taken to improve efficiency. To complete one update of the entire β vector, the program must compute the log likelihood function 70 times. The log likelihood function calculates $g(\mathbf{X}_i, \beta)$ given by Equation 1.3 for each i , where $i = 1, \dots, n$ and n is the number of rows in the data set. In order to reduce the number of iterations completed by the log likelihood function, a streamlined log likelihood function was called for the variables with the fewest number of ones. Such streamlining was possible because for variables composed of mostly zeroes, the only difference between the log likelihood calculation for the current versus the proposed β_i is the value of the proposed β_i . If the value of the variable is 0 the result of $g(\mathbf{X}_i, \beta)$ will be the same for the current and proposed values of β_i . Hence, those rows can be excluded from the calculation. Computer memory constraints allowed for this more efficient log likelihood calculation to be applied to the estimation of 30 of the 70 coefficients.

This program was tested on two differed Unix servers. The difference in run time between the two servers was significant. The university’s server took at least 10 times as long to update the β vector as compared to the server owned by the insurance company. This improvement was observed despite the fact that the insurance company’s server only allows a process to use a maximum of 25% of its CPU.

Ideally, we would have liked to run the program for a longer time period to better

ascertain the convergence of the Markov Chain Monte Carlo algorithm. Time constraints only allowed for the completion of 12,112 iterations. The first 7,609 estimates were used for burn in and are excluded from all parameter inference calculations.

4.3 Future Analyses

Based on the results of this model, our analysis could continue in many different directions. The ideas mentioned here are certainly not inclusive but could provide further insight into the underlying causes of the insurer's current enrollment and profitability issues. Separate models could be built for different segments of the population. For instance, a model could be built to determine the variables that influence disenrollment at the client level rather than the individual level. Another model could be constructed to determine the factors that influence a member's choice of product in an effort to predict movement between products offered by the same insurer. Yet another study could be conducted to determine whether the size of the client, estimated here by **NumProducts**, influences member disenrollment.

Another area for future work lies in improving the computational efficiency of this and other similar models that analyze large amounts of data. Increasing iteration speed and decreasing time to convergence are desirable improvements to the current model.

References

Draper, D. A., J. F. Hoadley, J. Mittler, S. Kuo, G. J. Bazzoli, P. J. Cunningham, and L. M. Nichols, L. J. Conwell. 2003. "Health care cost concerns intensify in Little Rock". Center for Studying Health System Change, Community Report no. 8.

Gilks, W. R., S. Richardson, and D.J. Spiegelhalter. 1996. *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Graves, T.L. and J.C. Kern. 2003. "Software testing and reliability modelling for Army Systems." Los Alamos National Laboratory Report LA-UR-03-0077.

Healthinsurance.info. 2002-2003. "Adverse selection and cream skimming."

Sutton, H., R. Feldman, and B. Dowd. 2002. "Disruption of a managed competition environment by low-ball premium bids: The Minnesota state employees group insurance program." Division of Health Services Research and Policy, University of Minnesota School of Public Health.

Appendix A

Initial Values: $\beta_i^{(0)}$

i	Estimate	i	Estimate	i	Estimate	i	Estimate	i	Estimate
0	0.0171	14	0.1501	28	0.0611	42	0.0096	56	0.7371
1	0.0270	15	0.0092	29	0.0599	43	0.0222	57	0.0566
2	0.0002	16	0.0089	30	0.1113	44	0.1699	58	0.4598
3	0.0001	17	0.0143	31	0.0111	45	1.4496	59	0.0783
4	0.0140	18	0.0457	32	0.0415	46	0.1560	60	0.0223
5	0.0917	19	0.0054	33	0.7303	47	0.0531	61	0.2474
6	0.0272	20	0.0086	34	0.0462	48	0.0864	62	0.0752
7	0.0826	21	0.0099	35	0.0591	49	0.3443	63	0.2275
8	0.0551	22	0.0169	36	0.0550	50	0.2973	64	0.0342
9	0.6994	23	0.0183	37	0.0284	51	0.3673	65	0.2161
10	0.5651	24	0.1129	38	0.4568	52	0.1677	66	0.0829
11	0.0087	25	0.1253	39	0.0414	53	0.5433	67	0.2772
12	0.8084	26	0.0213	40	0.0913	54	0.0463	68	0.7515
13	0.0611	27	0.0026	41	0.5346	55	0.0974	69	0.0688

Appendix B

Expert Values

i	q_i	i	q_i	i	q_i	i	q_i	i	q_i
1	0.9	15	0.7	29	0.3	43	0.5	57	0.7
2	0.7	16	0.5	30	0.3	44	0.3	58	0.3
3	0.9	17	0.5	31	0.5	45	0.9	59	0.3
4	0.7	18	0.5	32	0.5	46	0.3	60	0.5
5	0.7	19	0.7	33	0.3	47	0.5	61	0.3
6	0.7	20	0.5	34	0.7	48	0.3	62	0.5
7	0.3	21	0.5	35	0.3	49	0.5	63	0.3
8	0.7	22	0.7	36	0.3	50	0.3	64	0.5
9	0.9	23	0.5	37	0.5	51	0.3	65	0.5
10	0.9	24	0.3	38	0.3	52	0.3	66	0.3
11	0.5	25	0.5	39	0.7	53	0.5	67	0.3
12	0.7	26	0.5	40	0.3	54	0.3	68	0.3
13	0.5	27	0.3	41	0.3	55	0.3	69	0.3
14	0.9	28	0.5	42	0.3	56	0.3		

Appendix C

Table of Estimates

Variable Name	Mean Value	Proportion of Estimates Equal to 0
Intercept	1.69211	0
Prospective HealthRisk	-0.01234	0
Copay	-0.00006	0
Number of Months Eligible	-0.11943	0
Annual Expenditures	-0.00001	0
Age	-0.01198	0
Product	-0.01968	0
Number of Products Offered	-0.12816	0
Family	0.00018	0.806
Regional	0.00012	0.764
Risk	0.00877	0.002
Male	0.00142	0.530
Child	0.00009	0.812
No Diagnoses	0.00620	0.008
Spouse	0.00408	0.268
Client Cancelled	1.90995	0
ACC's:		
Ears Nose & Throat	0.00052	0.691
Musculoskeletal	0.00385	0.304
Skin Related	0.00384	0.309
Metabolic	0.00380	0.302
Heart	0.00369	0.345
Gastrointestinal	0.00376	0.311
Genital System	0.00406	0.292
Lung	0.00392	0.294
Infectious & Parasitic	0.00356	0.318
Mental Disorder	0.14339	0
Eye	0.00390	0.311

Variable Name	Mean Value	Proportion of Estimates Equal to 0
ACC's (continued)		
Uncertain Neoplasm	0.00391	0.286
Neurological	0.00398	0.306
Urinary System	0.00379	0.298
Diabetes	0.00417	0.261
Hematological	0.00388	0.303
Malignant Neoplasm	0.00388	0.265
Pregnancy Related	0.00168	0.013
Vascular Disease	1.14930	0
Substance Abuse	1.01289	0
Neonates	0.19817	0.050
Liver Disease	0.00395	0.289
Developmental Disorder	0.00377	0.326
Cerebro-Vascular Disease	0.00390	0.304
Cognitive Disorder	0.00386	0.309
Cardio-Respiratory	1.15581	0
Counties		
Westmoreland County	0.00385	0.305
Erie County	0.00401	0.289
Butler County	0.00383	0.312
Washington County	0.00404	0.285
Beaver County	0.82910	0
Cambria County	0.00374	0.312
Fayette County	0.00393	0.309
Blair County	0.00485	0.186
Armstrong County	0.00401	0.300
Mercer County	1.80536	0
Indiana County	0.00404	0.300
Lawrence County	0.09738	0
Somerset County	0.48232	0
Crawford County	0.00383	0.309
Jefferson County	0.00367	0.262
Clearfield County	0.00384	0.255
Elk County	0.00373	0.314
Clarion County	0.00409	0.290
McKean County	0.00384	0.274
Warren County	0.50781	0
Venango County	0.65666	0
Bedford County	0.00393	0.227
Greene County	0.00394	0.245
Huntingdon County	0.00384	0.243
Cameron County	0.00398	0.288
Potter County	0.00387	0.316
Centre County	0.00383	0.260
Forest County	0.00398	0.300

Appendix D

Actual Proportion of Members Who Left When $X_{ij} = 1$

Total Sample Proportion = 0.2194.

Variable Name	Proportion
Client Cancelled	0.5620
Beaver County	0.3151
Centre County	0.3067
Warren County	0.2918
Cardio-Respiratory	0.2818
Substance Abuse	0.2777
Mercer County	0.2773
Blair County	0.2710
No Diagnoses	0.2684
Pregnancy Related	0.2589
Venango County	0.2526
Risk Rated	0.2469
Cognitive Disorder	0.2423
Jefferson County	0.2420
Clearfield County	0.2361
Mental Disorder	0.2268
Developmental Disorder	0.2197
Male	0.2195
Neonates	0.2153
Liver Disease	0.2142
Child	0.2137
Urinary System	0.2137

Variable Name	Proportion
Family Contract	0.2135
Lawrence County	0.2131
Elk County	0.2129
Cerebro-Vascular Disease	0.2099
Neurological	0.2089
Forest County	0.2086
Vascular Disease	0.2081
Lung	0.2050
Spouse	0.2040
Infectious & Parasitic	0.2036
Somerset County	0.2034
Hematological	0.2031
Genital System	0.2028
Regional Account	0.2021
Fayette County	0.2017
Cameron County	0.2015
Huntingdon County	0.2013
Ears Nose & Throat	0.2008
Gastrointestinal	0.1999
Skin Related	0.1960
Musculoskeletal	0.1934
Eye	0.1932
Crawford County	0.1904
Butler County	0.1890
Erie County	0.1885
Diabetes	0.1883
McKean County	0.1882
Malignant Neoplasm	0.1878
Cambria County	0.1871
Clarion County	0.1845
Uncertain Neoplasm	0.1835
Bedford County	0.1834
Indiana County	0.1822
Metabolic	0.1790
Heart	0.1775
Westmoreland County	0.1717
Armstrong County	0.1678
Washington County	0.1675
Greene County	0.1387
Potter County	0.1377