베이지안통계특론I 기말대체과제

자동차사고 보험자료의 베이지안 추론 및 분석

202STG01 고유정

1. Introduction

General Insurance Association of Singapore에서 제공하는 자동차 보험 데이터에 ZIP 모형을 적 용해 보고자 한다. 이 자료에는 각 Policyholder에 대하여 policyholder의 성별, 나이, 사고건수, 차종, 자동차의 나이, 사고전적, 정책의 유효기간 등이 기록되어 있다. 전세계적으로 교통사고는 가장 빈번하게 일어나는 사고다. 교통사고 감소를 위한 대책을 세우기 위해서는 증가된 교통량과 관련된 교통사고 발생원인 및 특성을 규명하는 일이 우선되어야 한다. 교통사고의 원인별 규명이나 사전방지를 위해서는 정확한 사고통계자료를 바탕으로 한 사고모형 추정이 필수적이기에 자료수집과정에 대한 연구와 통계분석의 중요성이 강조된다. 본문에서는 교통사고분석에 이용되어온 통계분석기법을 관련응용사례 와 더불어 교통통계자료의 효과적인 활용을 위해 앞으로 연구되어야 할 부분과 적정보험료 산출에 대한 방향성 제시를 하고자 한다.

포아송 모델은 카운트 데이터를 모델링 하는데 가장 보편적으로 사용되는 모델이다. 그러나 현 데이터는 사고건수가 0이 과도하게 많은 영과잉 특징을 가진다. 즉, 포아송 모델이나 음이항 분포로 예측한 경우보다 0이 훨씬 많게 나온다. 특히 NCD(no claim discount system)로 인해 보험가입자들의 사고 보고율이 낮아지고 있다. 보험사에 보고되는 사고건수가 0이면 보험료(Premium)이 낮아지기 때문에 경미한 사고는 최대한 보험사에 연락하지 않고 개인적으로 해결하기 때문이다.

2. Method

본문에서는 자동차사고 보험 데이터에 ZIP 모형을 적용하여 베이지안 추론을 진행해보고자 한다. 먼저 Data 부문에서는 카운트 데이터의 변수를 설명하고 모델링에 적합한 변수들을 선택한다. 반응변수는 보고된 사고건수(Clm_Count)로 지정했고 설명변수는 14개의 변수 중 승용차 관련 변수들(차종, 승용차 나이, PC)과 Policyholder의 관한 변수들(성별, 나이, NCD)을 선택해 총6개의 변수를 설명변수로 지정했다. 그리고 JAGS를 이용해 영과잉 포아송 모형 ZIP을 사용하여 베이지안 추론과 분석을 수행해 보았다. 마지막으로는 모델링 결과 도출과 분석, 그리고 한계점에 대해 논하였다.

3. Data

General Insurance Association of Singapore에서 1993년에 발표한 자료로 싱가폴의 주요 보험사의 7,483개의 자동차보험 정책 자료다. 아래에 변수설명과 Claim counts(사고건수)에 대해 설명했으며 1년동안 보고된 최대 사고건수는 3이었다. 한 사람당 평균 사고건수는 0.06989로 0에 매우 가깝다. 이 분석의 목적은 자동차변수와 운전자 특성이 사고에 미치는 영향을 파악하기 위함이다. 따라서 영과잉 포아송 모델과 음이항 모델을 사용하여 더 적절한 분석모델을 찾고자 한다. 현 자료의 운전자(보험계약자) 중 90.64%가 남성이며 이를 통해 반응변수인 사고건수(Clm_Count)의 요인을 시각화로 추정해볼 수 있다. 본 분석에서는 Vehicle variable로 VAgeCat, Vehicle Type을, Policyholder(운전자) Variable로는 Female, AgeCat, NCD, PC를 사용하여 설명변수로 지정했다.

Table 1: 변수 설명

Variable	**Type**	**Explanation**
Clm_Count	Discrete	1년동안 사고건수(Claim Counts)
Female	Binary	Policyholder의 성별 0: 남성 1: 여성
AgeCat	Categorical	Policyholder(보험계약자)의 나이 0-6으로 그룹화되어 있으며 각각 21세 이하, 22-25세, 26-35세, 36-45세, 46-55세, 56-65세, 66세 이상을 나타낸다.
NCD	Categorical	No Claims Discount Policyholder의 사고전적 관련 변수로 무사고자들의 보험료 할인율이다. 0,10,20,30,40,50으로 이뤄져 할인율이 높을수록 사고전적이 좋은 편이다.
VAgeCat	Categorical	승용차의 나이 0-6으로 그룹화 되어있으며 각각 0, 1, 2, 3-5, 6-10, 11-15년을 뜻한다.
PC	Binary	1: 개인 승용차

		0: 그 외
Auto Age	Categorical	PC와 VAgeCat을 혼합해서 만든 범주형 변수
Vehicle Type	Categorical	보험이 적용되는 차종이며 A, G, M, P, Q, S, T, W, Z로 나타낸다.
Exp_weights	Numeric	정책의 유효 기간(년) 혹은 가중치
LNWEIGHT	Numeric	log(Exp_weights)

Fig.1 사고건수(Clm_Count)의 frequency 분포

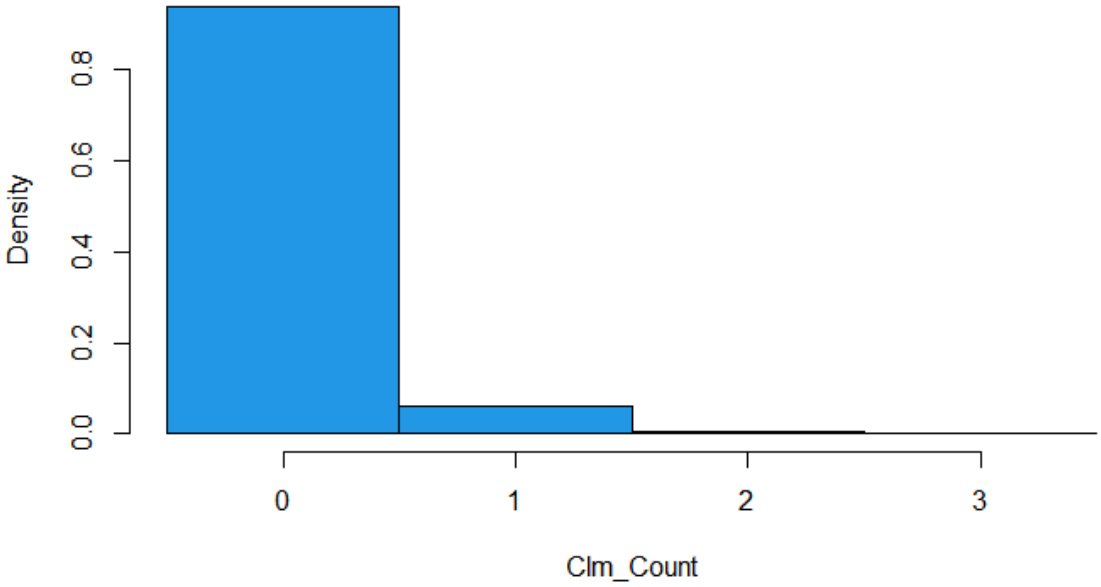

Table 2: 사고건수 frequency table

Claim count	0	1	2	3
Frequency	6996	455	28	4
% Claim count	93.49	6.08	0.37	0.05

Result & Analysis

영과잉 포아송 모델 ZIP 모델을 돌린 후 Gelman 상수를 보니 3개의 계수를 제외한 나머지 계수들에서 1에 매우 가까운 값을 가지므로 수렴이 이루어졌다고 볼 수 있다. Summary(codaSamples)를 해본 결과 로그선형 모형에서는 성별을 나타내는 'Female' 변수가, 로지스틱 연결함수에 대해서는 개인 자동차의 나이를 혼합하여 나타내는 'AgeCat1'이 제일 유의하다.

Fig.2

Potential scale reduction factors:

	Point est.	Upper C.I.
beta[1]	2.29	5.08
beta[2]	1.02	1.06
beta[3]	1.59	2.57
beta[4]	1.02	1.07
beta[5]	1.01	1.02
beta[6]	1.03	1.08
beta[7]	1.80	3.29
gamma[1]	3.25	6.38
gamma[2]	1.60	2.69
gamma[3]	3.56	6.72
gamma[4]	1.44	2.28
gamma[5]	1.34	1.91

Multivariate psrf

3.15

Fig.3

summary(codaSamples)

Iterations = 3200:13100
Thinning interval = 100
Number of chains = 3
Sample size per chain = 100

1. Empirical mean and standard deviation for each variable,
plus standard error of the mean:

	Mean	SD	Naive SE	Time-series SE
beta[1]	-1.51408	0.43490	0.025109	0.078392
beta[2]	-0.16030	0.15810	0.009128	0.008300
beta[3]	-0.09075	0.07053	0.004072	0.008485
beta[4]	0.02668	0.05930	0.003423	0.003749
beta[5]	-0.06876	0.10951	0.006322	0.010840
beta[6]	-0.14250	0.02987	0.001724	0.002008
beta[7]	-0.42455	0.34349	0.019832	0.054155
gamma[1]	-2.17483	2.21911	0.128121	0.574646
gamma[2]	-1.00773	0.53316	0.030782	0.140808
gamma[3]	1.24780	0.53257	0.030748	0.107906
gamma[4]	0.13740	0.23646	0.013652	0.040613
gamma[5]	-5.43598	2.98843	0.172537	0.628662

2. Quantiles for each variable:

	2.5%	25%	50%	75%	97.5%

```
beta[1]    -2.13460 -1.821324 -1.62692 -1.24865 -0.46661
beta[2]    -0.48436 -0.249909 -0.15535 -0.05076  0.15102
beta[3]    -0.23553 -0.131928 -0.08454 -0.04373  0.02539
beta[4]    -0.09287 -0.009885  0.02817  0.05918  0.14288
beta[5]    -0.28536 -0.146177 -0.06087  0.00018  0.13637
beta[6]    -0.20844 -0.159060 -0.14109 -0.12186 -0.08953
beta[7]    -1.22659 -0.630735 -0.38511 -0.17563  0.12736
gamma[1]   -6.69192 -3.776219 -2.06984 -0.43721  1.39152
gamma[2]   -2.13758 -1.312238 -0.92540 -0.59054 -0.17250
gamma[3]    0.38949  0.739773  1.31124  1.65390  2.23754
gamma[4]   -0.26065 -0.024201  0.10621  0.28479  0.69076
gamma[5] -11.83440 -7.302680 -5.02665 -2.90632 -1.17114
```

ZIP 모형에 대한 DIC를 30000번 iteration을 하여 돌렸을 때 아래와 같은 결과가 나온다. 음이항 분포의 Penalized deviance가 더 크므로 영과잉 포아송 모형이 더 적절하다고 판단할 수 있다.

- Mean deviance: 387.2
- Penalty: 264.4
- Penalized deviance: 611.8

또한 베타의 경로그림과 자기상관을 도출했을 때 충분히 수렴하고 자기상관이 낮아진 것을 확인할 수 있다. 그리고 표본으로 추정한 β, γ의 사후 밀도함수를 그려보면 정규분포에 가까운 형태를 띄고 있다. 제로 관측치에 대하여 관측된 0이 원조 제로로부터 유래할 확률을 그려보면 같은 제로 관측치라도 AgeCat1과 같은 자동차의 나이와 NCD(No Claim Discount)에 따라 원조제로 확률이 달라짐을 확인할 수 있었다.

Conclusion

앞에서 자동차 보험데이터의 효과적인 활용을 위해 적절한 분석기법을 찾아보았다. 데이터를 파악하고 시각화와 영과잉 포아송 모델을 통해 결과를 확인해 보았다.

모델링 결과 자동차 사고건수에 큰 영향을 끼치는 변수는 Policyholder의 성별, 차종, 그리고 자동차의 나이였다. 운전자의 90% 이상이 남자인 것을 보면 여자

에 비해 남성의 운전 속도가 빠르고 사고확률이 높다는 점을 판단 할 수 있다. 따라서, 사고율을 줄이기 위해 보험사에서 실시한 NCD의 효용성에 대해서 의문을 갖고 더 사고율을 효과적으로 낮추는, 즉, NCD처럼 오용의 확률이 낮은 전략을 세울 필요가 있다. 자동차의 나이 또한 운전자들이 개인적으로 고려해서 운전을 하는 것이 안전운전을 하기 위해 중요한 요소임을 강조해서 알릴 필요가 있다.

본 분석에서 아쉬웠던 점은 예상보다 Gelman 상수가 크게 나오는 계수들이 있었으며 유의한 계수가 많지 않았다는 점이다. 먼저 시간적 여유가 더 있어서 iteration을 더 크게 늘려서 돌렸다면 더 1에 가까운 Gelman 상수가 나왔을 것이다. 둘째, glm을 돌릴 때 교호작용을 고려하여 log(Exp_Weights)변수를 offset으로 놓고 모델링을 진행했다면 유의한 계수들이 많이 나왔을 것이라고 기대한다.

결과적으로 처음에 예상했던 대로 영과잉 카운트 데이터에 포아송 회귀모형을 적용하여 모델링 했을 때 적절하고 적합한 베이지안 분석기법임을 확인했다. 추후 여러 보험사에서 본문에서 제시한 방법과 방향성에 대해 깊이있는 이해와 연구 및 적용을 통하여 적절한 보험료 산출과 자동차 사고율을 낮추는 데에 기여하기를 기대한다.

Reference

1. Wagh, Y. S., & Kamalja, K. K. (2017). Modelling auto insurance claims in Singapore. *Sri Lankan Journal of Applied Statistics*, *18*(2), 105.
2. 손소영(1997). 교통사고 통계자료 분석 기법연구. 통계청『통계분석연구』제2권 제2호('97.가을호) 181-202
3. Jun, S. (2020). 베이지안 공액 사전분포를 이용한 키워드 데이터 분석. 한국콘텐츠학회논문지, 20(6), 1–8.
4. Faming Liang, Young K Truong & Wing Hung Wong. AUTOMATIC BAYESIAN MODEL AVERAGING FOR LINEAR REGRESSION AND APPLICATIONS IN BAYESIAN CURVE FITTING. Statistica Sinica 11(2001),

1005-1029.

5. 김명준, 김영화 (2009). 다양한 모형화를 통한 자동차 보험가격 산출. 한국데이터정보과학회지, 20(3), 515- 526

Appendix : 데이터 파일, R 코드

<R code>

```
library(insuranceData)

library(coda)

library(rjags)

library(runjags)


data(SingaporeAuto)

names(SingaporeAuto)

data <- SingaporeAuto


table(data$Clm_Count)

par(mfrow=c(1,1))

y=data$Clm_Count

hist(y, freq=FALSE, breaks=c(-0.5:(max(y)+1)),

     main=NULL, xlab='Clm_Count', col=4)

X=cbind(rep(1,length(y)), data$Female, factor(data$VehicleType),
factor(data$AgeCat), factor(data$VAgecat1), factor(data$NCD), data$PC)

Z=cbind(rep(1,length(y)), factor(data$VehicleType), factor(data$VAgecat1),
```

```
factor(data$NCD), data$PC)

p=ncol(X);q=ncol(Z)


modelString = "model{for(i in 1:length(y)){

y[i] ~ dpois(mu.pois[i])

mu.pois[i] <- (1-S[i])*lambda[i]+1e-10*S[i]

log(lambda[i]) <- inprod(X[i,], beta[])

S[i] ~ dbern(omega[i])

logit(omega[i]) <- inprod(Z[i,], gamma[])}

for ( i in 1:p){beta[i] ~ dnorm(mu.beta[i], Tau.beta[i])}

for ( i in 1:q){gamma[i] ~ dnorm(mu.gamma[i], Tau.gamma[i])}

}

"


writeLines(modelString, "model_ZIP_mixture.txt")


#prior parameters

mu.beta=rep(0,p)

Tau.beta=rep(0.01,p)

mu.gamma=rep(0,q)

Tau.gamma=rep(0.01,q)

glm.out=glm(y~X-1, family="poisson")
```

```
beta.pois=as.vector(glm.out$coefficients)

dataList=list(p=p, q=q, y=y, X=X, Z=Z, mu.beta=mu.beta, Tau.beta=Tau.beta,mu.gamma=mu.gamma, Tau.gamma=Tau.gamma)

initsList=list(beta=beta.pois, gamma=mu.gamma)

require(rjags);

jagsModel.zip=jags.model(file="model_ZIP_mixture.txt", data=dataList, inits=initsList, n.chains=3, n.adapt=100)

update(jagsModel.zip, n.iter=3000)

codaSamples=coda.samples(jagsModel.zip, variable.names=c("beta","gamma"), thin=1, n.chains=3, n.iter=10000, nthin=10)

coda::gelman.diag(codaSamples)

Summary(codaSamples)

dic.zip = dic.samples(jagsModel.zip, 30000); dic.zip
```