이론통계학2-Project #6 GLM을 이용한 자동차보험료 계산

강아미, 고유정, 윤보인, 이혜린, 홍지원

방법:

1) 각 팀별로 아래 1977년 스웨덴 자동차 3자 보험 자료를 분석
2) 팀별 분석결과를 A4지 10p 이내로 요약, 정리하여 e-mail로 제출
3) 분석내용 및 개발한 자동차보험료 계산 Application을 수업시간에 조별 발표

Part1) 사고빈도 GLM Model

a) 각 수준조합별 연간 사고빈도 확률변수 N_{ijkl}가 포아송 분포를 따른다고 가정할 때 적절한 최적 회귀모형을 구축하시오.

$$\ln \lambda_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + (\alpha\beta)_{ij} + \cdots$$

	AIC	BIC
Simple Model	10653.42	10970.52
MODEL 1	10526.39	11452.32
MODEL 2	10607.31	11228.82
MODEL 3	10505.87	11127.38
MODEL 4	10613.15	11386.87
MODEL 5	10382.66	11613.00
Full Model	10243.59	13249.69

Simple Model: 1차항만을 포함한 모형
Model1: 1차항과 Make, Bonus의 교호작용을 포함한 모형
Model2: 1차항과 Kilometres, Zone의 교호작용을 포함한 모형
Model3: 1차항과 Kilometres, Bonus의 교호작용을 포함한 모형
Model4: 1차항과 Zone, Bonus의 교호작용을 포함한 모형
Model5: 1차항과 Make, Bonus의 교호작용을 포함한 모형
Full Model: 모든 교호작용을 포함한 모형
AIC를 기준으로 선택한 모형: Full Model
BIC를 기준으로 선택한 모형: Simple Model

[BIC]

Variable	Coefficient	Variable	Coefficient
(Intercept)	-2.2596259	Bonus.C	-0.2369782
Kilometres.L	0.42494642	Bonus^4	-0.0208448
Kilometres.Q	-0.0282943	Bonus^5	-0.0471208
Kilometres.C	0.06065036	Bonus^6	-0.0073805
Kilometres^4	0.00336642	Make2	0.07621416
Zone2	-0.2381499	Make3	-0.2472338
Zone3	-0.3863485	Make4	-0.6536623
Zone4	-0.5818801	Make5	0.15493887
Zone5	-0.3259656	Make6	-0.3356108

Zone6	-0.5262435	Make7	-0.0563043
Zone7	-0.7322132	Make8	-0.0439725
Bonus.L	-0.9909745	Make9	-0.0680777
Bonus.Q	0.16692001		

b) 각 수준조합별 연간 사고빈도 확률변수 N_{ijkl}이 이항분포를 한다고 가정 할 때 적절한 최적 회귀모형을 구축하시오.

	AIC	BIC
Simple Model	10783.41	11100.51
MODEL 1	10607.78	11533.71
MODEL 2	10734.81	11356.32
MODEL 3	10626.49	11248.01
MODEL 4	10671.74	11445.46
MODEL 5	10460.42	11690.76
Full Model	10239.63	13245.73

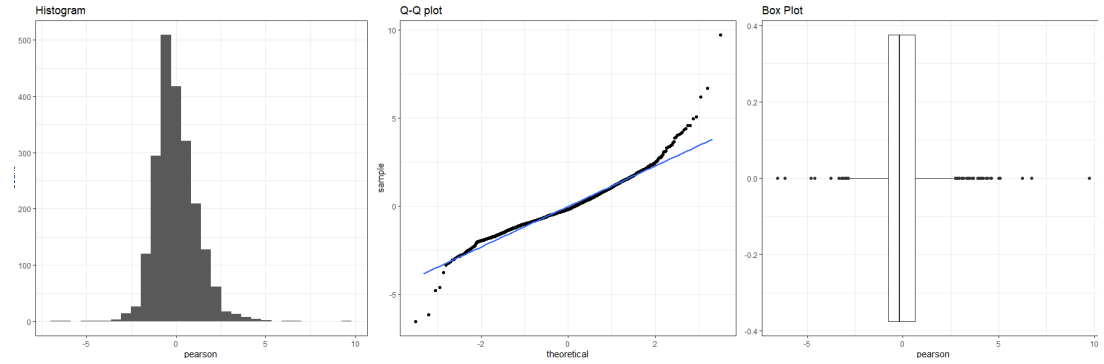
Simple Model: 1차항만을 포함한 모형
Model1: 1차항과 Make, Bonus의 교호작용을 포함한 모형
Model2: 1차항과 Kilometres, Zone의 교호작용을 포함한 모형
Model3: 1차항과 Kilometres, Bonus의 교호작용을 포함한 모형
Model4: 1차항과 Zone, Bonus의 교호작용을 포함한 모형
Model5: 1차항과 Make, Bonus의 교호작용을 포함한 모형
Full Model: 모든 교호작용을 포함한 모형
AIC를 기준으로 선택한 모형: Full Model
BIC를 기준으로 선택한 모형: Simple Model

[BIC]

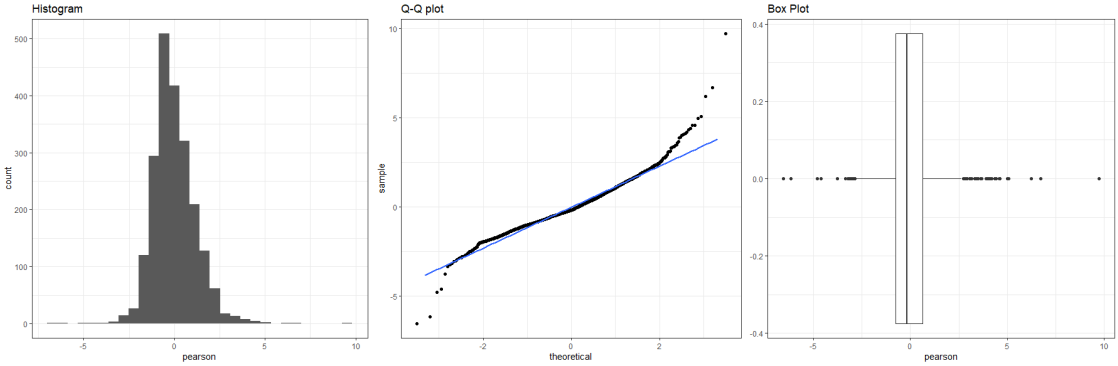
Variable	Coefficient	Variable	Coefficient
(Intercept)	-2.1482851	Bonus.C	-0.2588907
Kilometres.L	0.45477869	Bonus^4	-0.0161999
Kilometres.Q	-0.0287198	Bonus^5	-0.0510852
Kilometres.C	0.06599011	Bonus^6	-0.0072021
Kilometres^4	0.00393475	Make2	0.08173063
Zone2	-0.2597665	Make3	-0.261787
Zone3	-0.4180836	Make4	-0.6960815
Zone4	-0.6245381	Make5	0.16708021
Zone5	-0.3539251	Make6	-0.3560686
Zone6	-0.5664218	Make7	-0.0600081
Zone7	-0.78213	Make8	-0.0477201
Bonus.L	-1.0685577	Make9	-0.0723895
Bonus.Q	0.19647662		

c) 위의 두 GLM모형에서 추정한 평균사고빈도 $\hat{\lambda}_{ijkl}, \hat{p}_{ijkl}$ 값과 단순추정값 $\bar{\lambda}_{ijkl} = n_{ijkl}/m_{ijkl}$을 이용한 Pearson 표준화 잔차 및 Deviance 잔차를 각 수준별로 구하고 이들 잔차값들의 boxplot, histogram, normal Q-Q plot을 그려보고 추정모형의 적합도를 검토하시오.

Pearson Residual



Deviance Residual



Pearson 표준화 잔차와 Deviance 잔차로 Histogram 을 그려본 결과 0 을 중심으로 고르게 분포한 것을 확인할 수 있습니다. 뿐만 아니라, Q-Q plot 역시 오른쪽 꼬리를 제외하고 대부분의 구간에서 직선 모형을 띄므로 정규성을 따른다고 볼 수 있습니다.

d) a, b 에서 추정한 두 GLM 모형의 차이점 및 장단점을 비교 검토하시오.

⇨ 총 사고 건수(y)가 포아송분포를 따른다고 가정시 도출한 회귀모형의 경우 , y값에는 로그를 취하므로 y값이 0보다 큰 값에 대해서만 추정할 수 있다.

⇨ 총 사고 건수(y)가 이항분포를 따른다고 가정시 도출한 회귀모형의 경우,
총 사고건수와 보험가입자수 – 총 사고건수 모두 0이상인 값에 대해서만 추정할 수 있다.

⇨ 포아송분포를 가정한 모형의 예측력이 높았다. [1977-스웨덴자동차 보험_1956-캐나다자동차보험자료]자료가 포아송분포에 적절했음을 알 수 있다.

e) 평균사고빈도에 대한 단순 추정값과 GLM을 이용한 추정량의 장단점을 검토하시오.

⇨ 평균 사고 빈도를 단순 평균으로 추정하는 경우, 자료가 없는 변수들의 조합에 대해서는 사고 빈도를 추정할 수 없기 때문에 전체 데이터에 대한 평균 사고 빈도의 값으로 추정해야 한다.
반면 GLM 모형을 사용하면 자료에 포함되어 있지 않은 수준 조합에 대해서도 사고 빈도를 추정할 수 있다는 장점이 있다.

f) 사고빈도의 분포로 포아송분포 대신 음이항분포를 이용할 경우 차이점 및 장단점을 제시하시오.

	AIC	BIC
Simple Model	10371.89	10688.99
MODEL 1	10320.82	11246.75
MODEL 2	10388.05	11009.56
MODEL 3	10338.18	10959.69
MODEL 4	10395.34	11169.06
MODEL 5	10278.57	11508.91
Full Model	10238.01	13244.11

Simple Model: 1차항만을 포함한 모형

Model1: 1차항과 Make, Bonus의 교호작용을 포함한 모형

Model2: 1차항과 Kilometres, Zone의 교호작용을 포함한 모형

Model3: 1차항과 Kilometres, Bonus의 교호작용을 포함한 모형

Model4: 1차항과 Zone, Bonus의 교호작용을 포함한 모형

Model5: 1차항과 Make, Bonus의 교호작용을 포함한 모형

Full Model: 모든 교호작용을 포함한 모형

AIC를 기준으로 선택한 모형: Full Model

BIC를 기준으로 선택한 모형: Simple Model

Variable	Coefficient	Variable	Coefficient
(Intercept)	-2.2559849	Bonus.C	-0.2192373
Kilometres.L	0.37836467	Bonus^4	-0.0361596
Kilometres.Q	-0.0154883	Bonus^5	-0.0346173
Kilometres.C	0.05996304	Bonus^6	-0.0095915
Kilometres^4	0.00064779	Make2	0.06720853
Zone2	-0.2241638	Make3	-0.2349482
Zone3	-0.3827039	Make4	-0.6838033
Zone4	-0.5558029	Make5	0.15227488
Zone5	-0.3383065	Make6	-0.3633951
Zone6	-0.5224938	Make7	-0.0796068
Zone7	-0.7321634	Make8	-0.0412292
Bonus.L	-1.0049032	Make9	-0.0902704
Bonus.Q	0.15849127		

⇨ 포아송분포, 이항분포, 음이항분포 모두 AIC를 기준으로 한 경우 모든 교호작용을 포함한 모형이 선택되었습니다. 반면 BIC를 기준으로 한 경우 1차항만을 포함한 모형이 선택되었습니다. 따라서 모든 결과가 일관성 있음을 확인할 수 있습니다.

Part2) 사고심도 GLM Model

a) 사고 1건당 보험금 확률변수 y_{ijkl}가 Gamma 분포를 따른다고 할 때 적절한 최적회귀 모형을 구축하시오.

	AIC	BIC	F-test
simple Model	1878028	1878269	0.03342*
Model1	**1873965**	**1874553**	**0.1756**
Model1-2	1879000	1878908	0.01661*
Model2	1876010	1876482	0.04632*
Model3	1876389	1876861	0.02774*
Model4	1875732	1876281	0.03*
Full Model	1859773	1862057	

simple model : 1차항만 포함한 모델

model 1 : simple model에서 Zone*Bonus 항 추가

model 1-2 : Bonus, Kilometres를 numeric 변수로 지정한 모델

model 2 : simple model에서 Zone*Kilometres 항 추가

model 3 : simple model에서 Bonus*Kilometres 항 추가

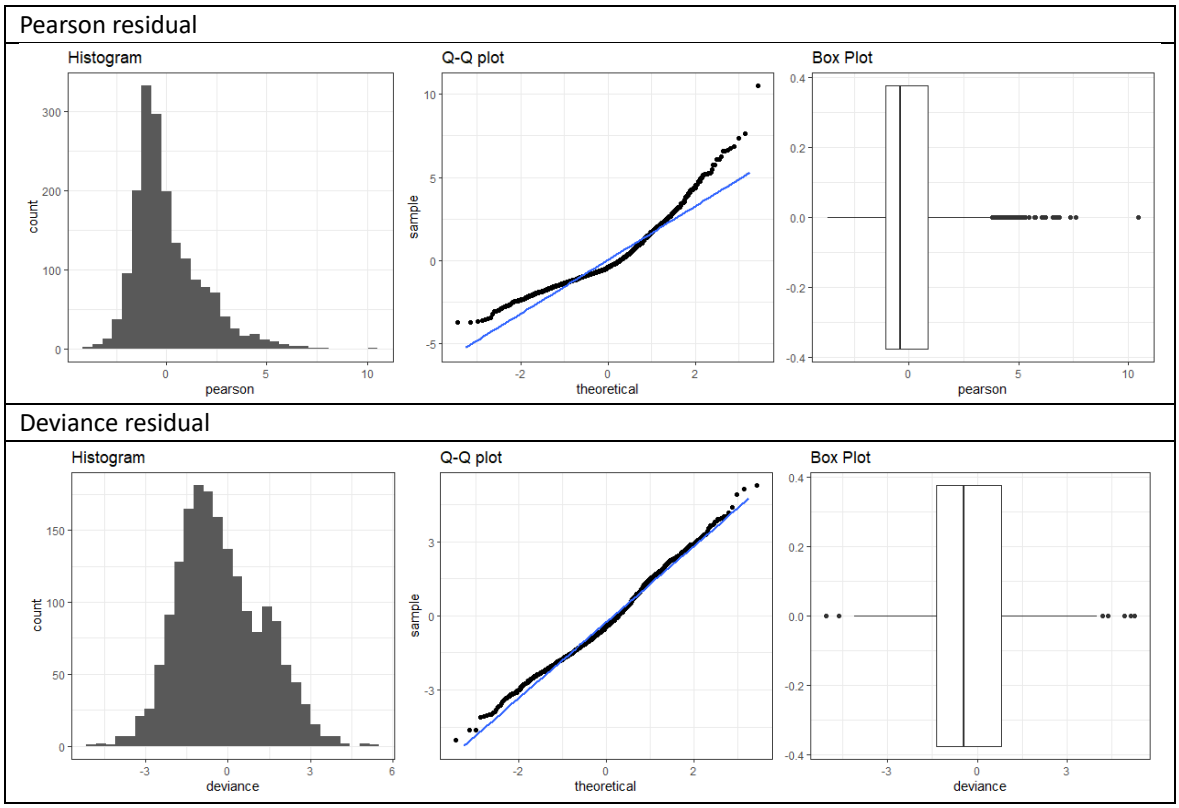
model 4 : simple model에서 Make*Kilometres 항 추가

full model : 모든 교호작용 포함한 모델

AIC, BIC 기준으로 모든 교호작용이 포함된 full model이 가장 유의했다. 모형의 복잡성을 완화하기 위해 교호작용 하나를 포함한 모형과 full model의 F-test를 시행해본 결과, model1과의 F-test 결과에서 [H0: 추가항의 계수가 0이다.] 의 가설을 기각할 수 없다는 결론이 나오므로 model1을 최적의 모형으로 결론내린다.

Variable	Coefficient	Variable	Coefficient	Variable	Coefficient	Variable	Coefficient
(Intercept)	8.480	Bonus7	0.006	Zone2:Bonus3	0.073	Zone5:Bonus5	0.021
Kilometres2	0.031*	Make2	-0.034	Zone3:Bonus3	0.018	Zone6:Bonus5	0.046
Kilometres3	-0.008	Make3	0.085*	Zone4:Bonus3	0.013	Zone7:Bonus5	-0.370
Kilometres4	0.001	Make4	-0.164***	Zone5:Bonus3	-0.009	Zone2:Bonus6	0.046
Kilometres5	-0.011	Make5	-0.089*	Zone6:Bonus3	0.019	Zone3:Bonus6	0.058
Zone2	0.006	Make6	-0.041	Zone7:Bonus3	0.424	Zone4:Bonus6	0.076
Zone3	0.037	Make7	-0.119**	Zone2:Bonus4	-0.099*	Zone5:Bonus6	0.321***
Zone4	0.131***	Make8	0.213***	Zone3:Bonus4	-0.095	Zone6:Bonus6	0.112
Zone5	0.053	Make9	-0.055**	Zone4:Bonus4	-0.096*	Zone7:Bonus6	0.255
Zone6	0.147***	Zone2:Bonus2	0.021	Zone5:Bonus4	-0.214**	Zone2:Bonus7	-0.121
Zone7	-0.105	Zone3:Bonus2	0.083*	Zone6:Bonus4	-0.205**	Zone3:Bonus7	-0.040
Bonus2	0.030	Zone4:Bonus2	0.036	Zone7:Bonus4	-0.066	Zone4:Bonus7	0.057
Bonus3	-0.015	Zone5:Bonus2	0.082	Zone2:Bonus5	0.083	Zone5:Bonus7	-0.017
Bonus4	0.141 ***	Zone6:Bonus2	0.074	Zone3:Bonus5	0.027	Zone6:Bonus7	0.045
Bonus5	-0.016*	Zone7:Bonus2	0.130	Zone4:Bonus5	-0.034	Zone7:Bonus7	-0.334
Bonus6	-0.086						

b) GLM 모형에서 추정한 평균사고심도 $\hat{\mu}_{ijkl}$값 단순추정값 $\bar{y}_{ijkl} = y_{ijl}/m_{ijkl}$을 이용한 Pearson 표준화 잔차 및 Deviance 잔차를 각 수준별로 구하고 이들 잔차들의 boxplot, histogram, normal Q-Q plot을 그려보고 추정 모형의 적합도를 검토하시오.



Deviance 잔차가 정규성을 잘 따르는 것으로 확인된다. Deviance 잔차의 경우 양쪽 꼬리를 제외하고는 Q-Q plot이 직선을 잘 따르는 것으로 보인다.

c) 평균사고심도에 대한 단순추정값과 GLM을 이용한 추정량의 장단점을 검토하시오.
(예를 들어 사고건수가 작거나 0인 경우 차이점 검토)

Kilometres	Zone	Bonus	Make	Claims	Payment	fit
3	7	3	4	0	0	2413.402
1	7	3	7	0	0	2468.929
3	7	3	7	0	0	2523.094
1	7	3	5	0	0	2545.36
5	7	3	7	0	0	2568.537
3	7	3	5	0	0	2601.202
...	...	...	...	...	...	...
1	5	5	8	0	0	7347.699
3	7	2	8	0	0	7408.037
2	7	2	8	0	0	7425.244
5	6	5	8	0	0	7637.625

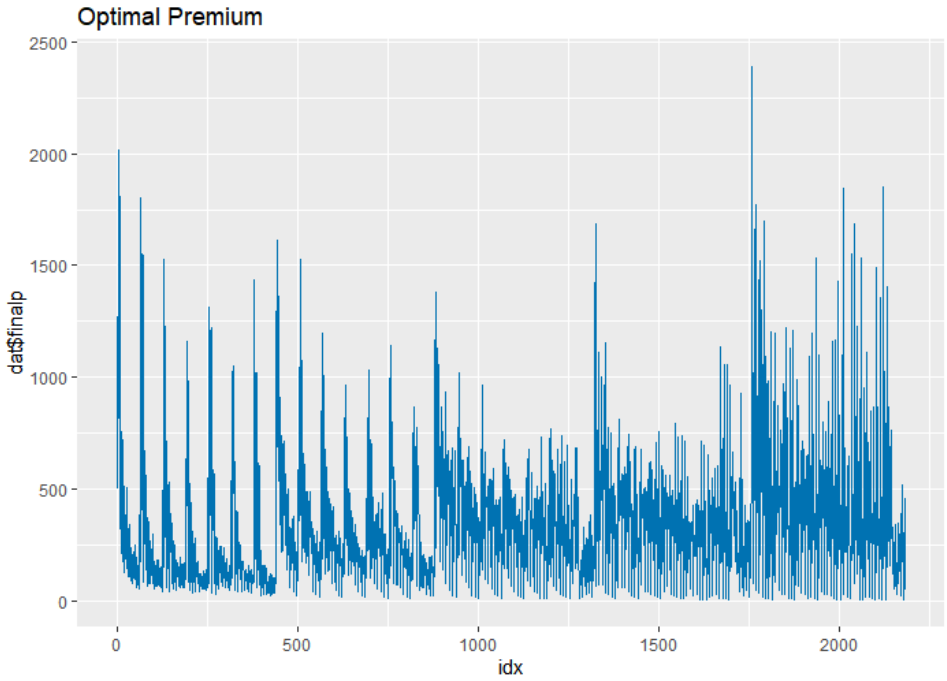
5	5	5	8	0	0	7644.139
4	5	5	8	0	0	7663.647

평균 사고심도를 단순 평균으로 추정하면, 자료가 없는 변수들의 조합에 대해서는 사고 심도를 추정할 수 없기 때문에 전체 데이터에 대한 평균 사고심도의 값(4955.251)으로 추정해야 한다. 하지만 GLM 모형을 사용하면 자료에 포함되어 있지 않은 수준 조합에 대해서도 사고 심도를 추정할 수 있다는 장점이 있다.

Part3) 할증 보험료 자동계산

a) 위의 두 회귀모형을 이용하여 보험 가입자 및 차량유형(주행거리, 운전지역, 무사고 보너스, 차종)별로 차등한 된 x년간 자동차 보험료를 산출하는 공식을 제시하시오.

	GLM
1인당 년간 평균사고빈도	$\widehat{\lambda_{ijkl}} = n_{ijkl}/m_{ijkl}$ = 총사고건수 / 보험가입자수 **추정된 회귀식 :** $\ln\widehat{\lambda_{ijkl}} = \ln n_{ijkl}/m_{ijkl}$ $= \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + (\alpha\beta)_{ij} + \cdots$
1건당 평균사고심도	$\widehat{\mu_{ijkl}} = y_{ijkl}/n_{ijkl}$ = 총보험금 지불액 / 총사고건수 **추정된 회귀식** : $\ln\widehat{\mu_{ijkl}} = \ln y_{ijkl}/n_{ijkl}$ $= \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + (\alpha\beta)_{ij} + \cdots$
수준조합별 적정보험료	$\widehat{p_{ijkl}}$ = 1인당 년간 평균사고빈도* 1건당 평균사고심도 = $\widehat{\lambda_{ijkl}} * \widehat{\mu_{ijkl}}$ 요인에 해당하는 변수 4가지 **주행거리(Kilometers)**, **운전지역(Zone), 무사고 보너스(Bonus), 자동차 종류(Make)**를 적용하고, 위 회귀분석을 통해 추정한 1인당 년간 평균사고빈도와 1건당 평균사고심도를 곱하여 교호작용을 고려한 각 수준조합별 적정보험료를 산출했다.

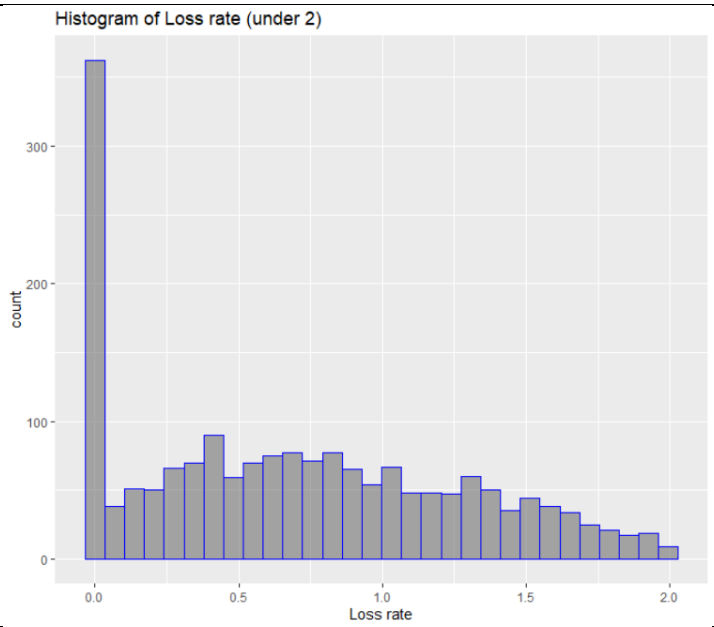


b) a)에서 GLM을 이용해 계산한 보험료를 적용했을 때 각 수준조합별로 손해율을 각각 구하고 이들 손해율 값의 boxplot 및 histogram을 그려서 수준별 보험요율의 적절성을 검토하시오.

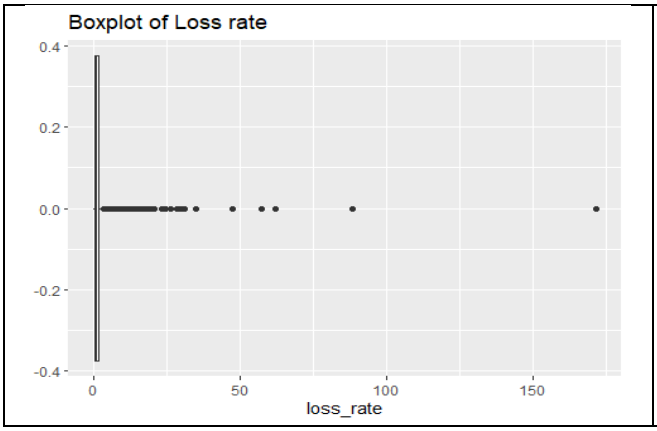
* 각 수준조합별로 구한 손해율 표

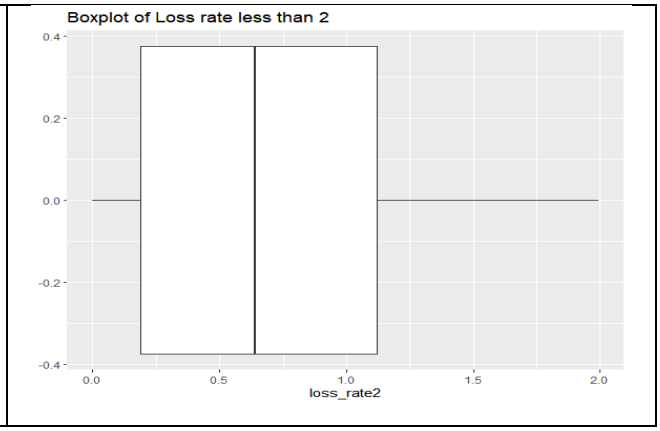
1.04	1.33	0.41	0.16	0.38	0.38	0.53	3.82	0.38	1.89	3.73
2.11	0.41	0.33	0.28	1.67	2.51	0.62	1.39	5.27	1.09	0.53
0.58	0.57	2.26	3.96	0.67	1.28	7.44	3.07	1.03	1.01	2.97
2.51	0.00	0.88	0.84	1.16	8.43	3.47	0.72	2.32	0.31	0.00
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

손해율 값의 Histogram



손해율의 Boxplot	손해율의 Boxplot





Min	1st Qu	Median	Mean	3rd Qu	Max	NA
0.00	0.27	0.79	1.62	1.45	171.47	25

2 미만 손해율의 요약통계량

Min	1st Qu	Median	Mean	3rd Qu	Max	NA
0.00	0.19	0.64	0.70	1.12	1.99	0

보험요율은 손해 발생 및 손해율에서 보험 원가를 부담할 순 보험료를 구한 후 예정 경비와 이윤을 가산하여 산출한다. 일반적으로 차등요율을 적용하기 때문에 위에 차등화 된 보험료를 구하여 수준조합별로 손해율을 구해보았다.

위에 히스토그램과 상자그림에서 볼 수 있듯이 손해율은 대부분 0 에서 2 사이에 집중적으로 포화되어 있다. 따라서 2 미만의 손해율값을 따로 추출하여 따로 박스플랏을 그려보았다.

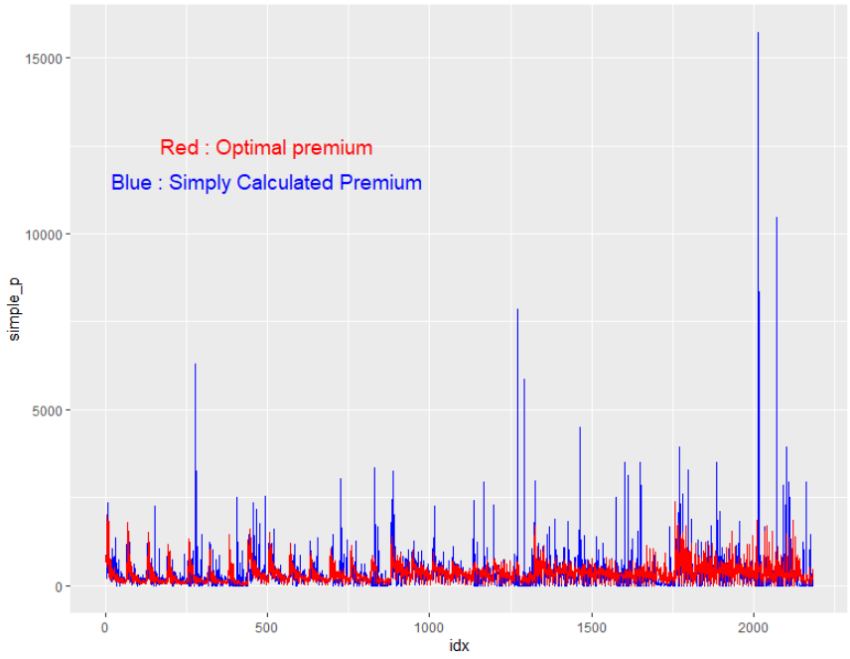
즉, 손해율이 0 에서 2 사이에 분포하여 높지 않고 큰 변동이 없기에 매우 안정적이라 볼 수 있다. 그러므로 현 보험요율은 적절하다고 할 수 있다.

c) 위에서 계산한 보험료를 실제 적용했을 때 전체손해율(총지급보험금 / 총보험료 = $\sum y_{ijkl} / [\sum p_{ijkl} m_{ijkl}]$)을 구하고 보험요율의 적정성을 검토하시오.

총보험료 = $\sum$추정한 보험료 $*$ 보험가입자수
전체손해율 Total loss rate = 총지급보험금 / 총보험료
= 0.9209751

전체손해율은 약 0.92로 1% 미만이며 정상적인 수치다. 따라서, b)처럼 손해율이 안정적임을 알 수 있다. 따라서 보험요율은 적절하다.

d) 각 수준별 단순계산보험료(y_{ijkl}/m_{ijkl} = 지불보험금 /보험가입자수)와 GLM을 이용한 추정보험료 p_{ijkl}의 차이점을 설명하고 두 방법의 장단점을 설명하시오.



1) 차이

위 그래프를 보면 대체로 단순계산보험료가 GLM을 이용하여 추정한 보험료보다 수치가 큰 것을 알 수 있다. 또한 이 차이는 급격하게 커지는 추세를 보인다.

2) GLM을 이용한 보험료 추정방법의 장단점

- 장점) 결측치와 오차를 줄이고 교호작용을 고려하고 1인당 평균사고빈도와 1건당 평균사고심도를 별도로 구하여 추정했기에 정확도가 높은 방법이다.
- 단점) 보험료 측정과 보험요율에 관한 전문적인 지식이 없는 대부분의 일반인들은 이해하기 어려운 방법이기에 쉽게 보편적으로 사용하기 어렵다.

3)단순계산을 통한 보험료 추정방법의 장단점

- 장점) 누구나 쉽게 이해하고 적용할 수 있다.
- 단점) 원 데이터의 지불보험금과 보험가입자 수만을 갖고 추정하였기에 정확도가 떨어진다.

e) 위의 part 1) 2)에서 추정된 GLM 회귀무형을 이용하여 건당 보상한도(limit) B 및 자기공제액(deductable) A가 있는 경우 각 보험가입자의 적정보험료 계산하는 Application을 개발하시오.

https://hyerin0113.shinyapps.io/final_hw6/