이론통계학2

Project #6. GLM을 이용한 자동차 보험료 계산

발표일 2021. 11. 03

1조

202STG26	박지윤
202STG27	이수현
212STG04	김이현
212STG12	박윤정

Part 1 : 사고 빈도 GLM Model

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

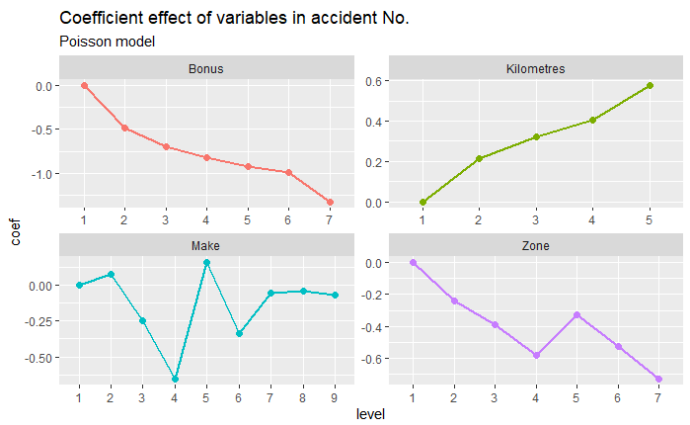
Bonus	Interaction	선택 기준	value	선택된 변수 개수
Catogorical	미포함	AIC	**10653.42**	25
		BIC	**10970.52**	25
	포함	AIC	10243.59	237
		BIC	12648.04	189
Numeric	미포함	AIC	11702.51	20
		BIC	11956.19	20
	포함	AIC	11235.50	142
		BIC	12436.32	94

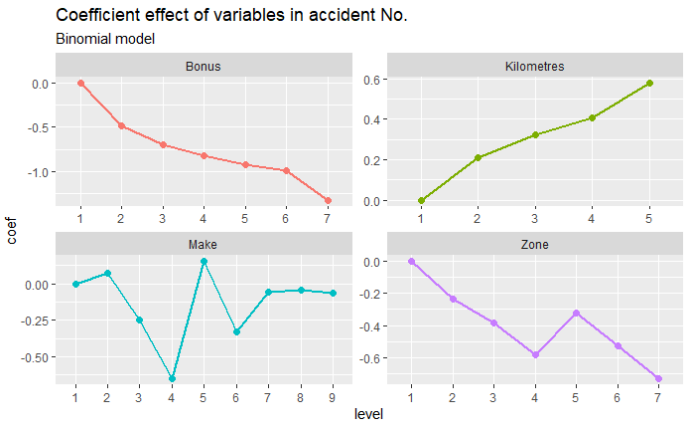
연간 사고빈도 확률변수가 포아송 분포 따른다고 가정한 후, 가입자수가 0보다 큰 자료에 대해 다양한 방법을 통해 최적 회귀 모형을 구하였고 결과는 위 표와 같다. Bonus 변수의 연속변수 간주 여부와 교호작용 효과의 포함 여부를 변화시켜 가며 모형을 적합하였고, AIC과 BIC를 기준으로 각각 모형을 선택하였다. 이때 AIC 기준으로는 bonus가 범주형 변수일 때, 교호작용 효과 포함 모형이 선택되었고 AIC 값은 10243.59였다. BIC 기준으로는 bonus가 범주형 변수일 때, 교호작용이 없을 때의 모형이 선택되었고, BIC값은 10970.52였다. 두 모델 중 1차 선형 모형이 교호작용을 포함한 모델보다 각 변수의 영향력을 파악하기 쉬우므로 BIC를 기준으로 선택한 모델을 최적 회귀 모형으로 선택한다.

2. 각 수준조합별 연간 사고빈도 확률변수 N_{ijkl}이 이항분포를 한다고 가정할 때 적절 한 최적 회귀모형을 구축하시오.

Bonus	Interaction	모형 선택 기준	value	선택된 변수 개수
Catogorical	미포함	AIC	10770.98	25
		BIC	**11088.08**	25
	포함	AIC	**10226.57**	237
		BIC	12634.64	189
Numeric	미포함	AIC	11972.20	20
		BIC	12225.88	20
	포함	AIC	11402.37	142
		BIC	12606.98	94

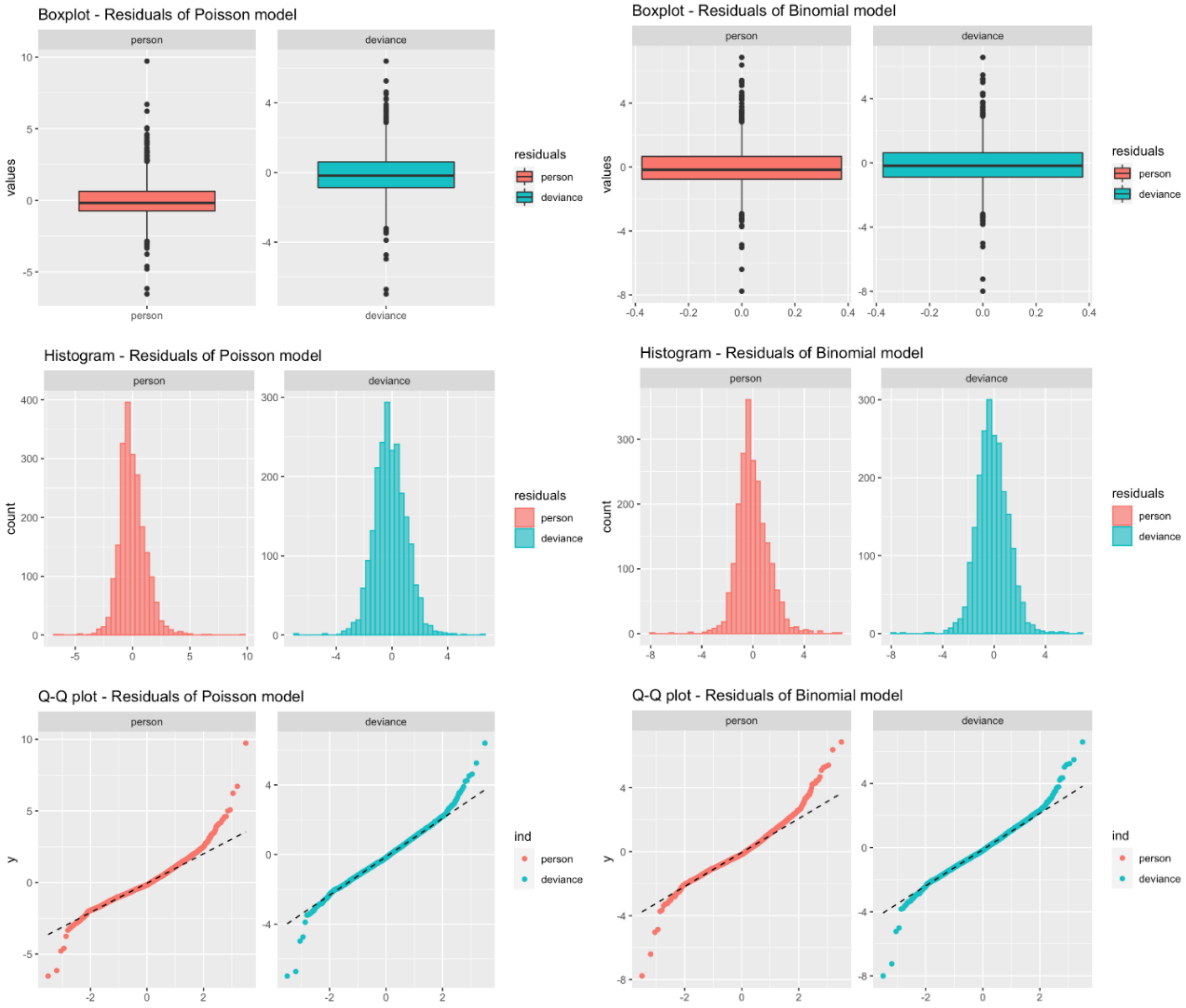
연간 사고빈도 확률변수가 이항 분포 따른다고 가정한 후, Bonus 변수의 연속변수 간주 여부, 교호작용 효과의 포함 여부를 변화시켜가며 AIC과 BIC를 기준으로 각각 모형을 선택하였다. 이때 AIC 기준으로는 bonus가 범주형 변수일 때 교호작용 효과를 포함한 모형이 선택되었고 AIC 값은 10226.57였다. BIC 기준으로는 bonus가 범주형 변수일 때 교호작용이 없는 모형이 선택되었고, BIC값은 11088.08였다. 1번에서와 마찬가지로 변수 효과 파악의 용이함을 위해 BIC를 기준으로 한 1차 선형 모형을 최적 회귀 모형으로 선택한다.





최적 예측 모형을 통해 구한 각 변수 별 효과의 그래프이다. 왼쪽은 포아송 분포를 사용한 모델, 오른쪽은 이항 분포를 사용한 모델이다. Bonus는 감소할수록, kilometres는 증가할수록 변수의 효과가 커짐을 파악할 수 있는데, 이는 무사고 기간이 짧을수록, 주행 거리가 길수록 λ가 커진다는 것을 뜻한다. 또한 make가 5일 때, zone이 1일 때 λ값이 큼을 알 수 있다.

3. **위의 두 GLM 모형에서 추정한 평균사고빈도 $\widehat{\lambda_{ijkl}}$, $\widehat{p_{ijkl}}$값과 단순추정값 $\overline{\lambda_{ijkl}}$을 이용한 Pearson 표준화 잔차 및 deviance 잔차를 각 수준별로 구하고 이들 잔차값들의 boxplot; histogram; normal-Q-Q plot을 그려보고 추정모형의 적합도를 검토하시오.**



위에서 선택한 두 가지의 최적 모형에 대해 pearson 표준화 잔차와 deviance 잔차를 구한 후 boxplot, histogram, normal q-q plot을 그려보았다. 왼쪽이 포아송 분포의 잔차, 오른쪽이 이항 분포의 잔차이고 적색이 pearson 잔차, 청색이 deviance 잔차이다. q-q plot에서 두 모형 모두 중간 부분은 잘 적합됐으나 양끝이 직선에서 벗어난다. 왼쪽 끝을 보면

포아송 모형이 이항 분포 모형보다 조금 더 잘 적합된 것으로 보인다.

4. 포아송 분포와 이항분포를 이용해 추정한 두 GLM 모형의 차이점 및 장단점을 비교 검토하시오.

두 모형 모두 적합도 면에서는 큰 차이를 보이지 않았다. 하지만 해당 데이터의 경우 총사고건수가 보험가입자수보다 많은 경우가 존재하였다. 이런 경우 이항 분포를 이용할 경우에는 해당 관측치를 제외하고 모형을 적합해야 했다. 이에 반해 포아송 분포를 이용한 모형은 이런 관측치에 영향을 받지 않아 데이터 전부를 이용하여 적합할 수 있었다.

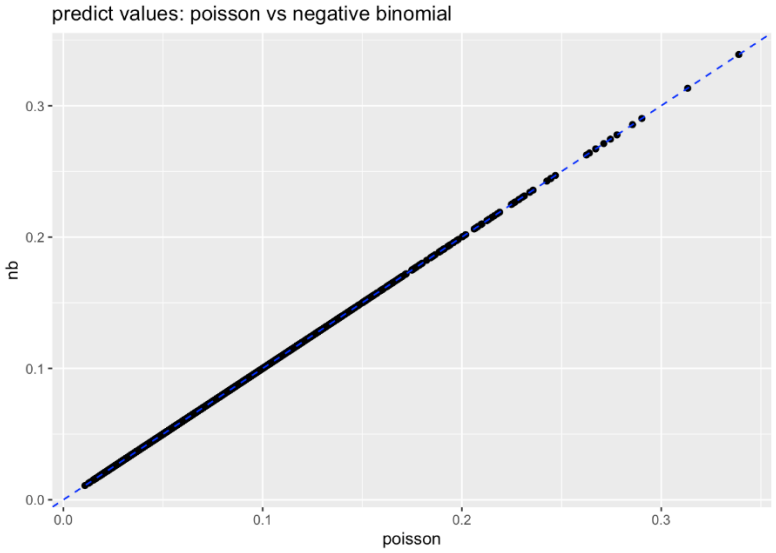
5. 평균사고빈도에 대한 단순 추정값과 GLM을 이용한 추정량의 장단점을 검토하시오.

보험가입자수	총사고건수	단순 추정값	GLM 추정값
15	2	0.133	0.097
18	3	0.167	0.078
13	0	0.000	0.068
18	0	0.000	0.062
0	0	NaN	0.038
0	0	NaN	0.028
0	0	NaN	0.047
0	0	NaN	0.104

보험 가입자수가 20보다 작은 경우에 대한 단순 추정값과 포아송 모형을 이용한 glm 모델로 구한 추정값들의 일부를 나타낸 표이다. 보험가입자수가 0인 경우 단순 추정값은 추정할 수 없지만 glm 모형의 경우 λ 추정이 가능하다. 또한 총 사고건수가 0인 경우에 대해서도 λ를 추정 가능해 보험료 산출이 가능하다.

6. 사고빈도의 분포로 포아송분포 대신 음이항분포(Negative Binomial)를 이용할 경우 차이점 및 장단점을 제시하시오.

음이항 분포를 사용했을 때의 교호 작용을 포함하지 않은 1차 선형 모형을 적합시킨 결과, theta = 738022인 모형이 선택되었고, 이때의 AIC값은 766157였다. Dispersion parameter가 아주 큰 값을 가지므로 포아송 분포를 이용했을 때의 모형과 비슷할 것임을 예상할 수 있다.



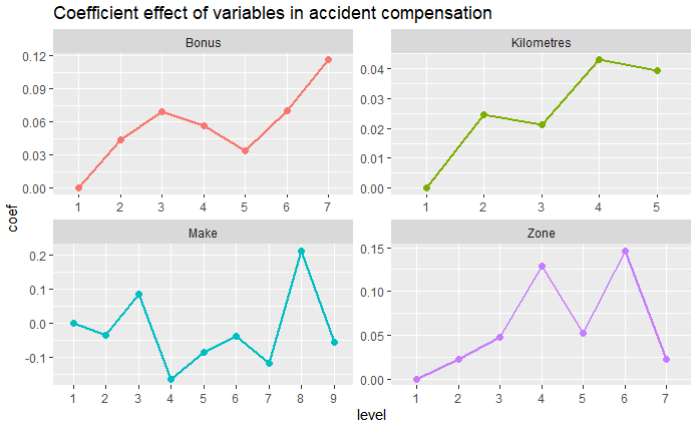
두 모형의 예측값을 비교한 그래프이다. 예상했던 것과 마찬가지로 두 예측값이 동일한 결과가 나왔다.

Part 2 : 사고 심도 GLM Model

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

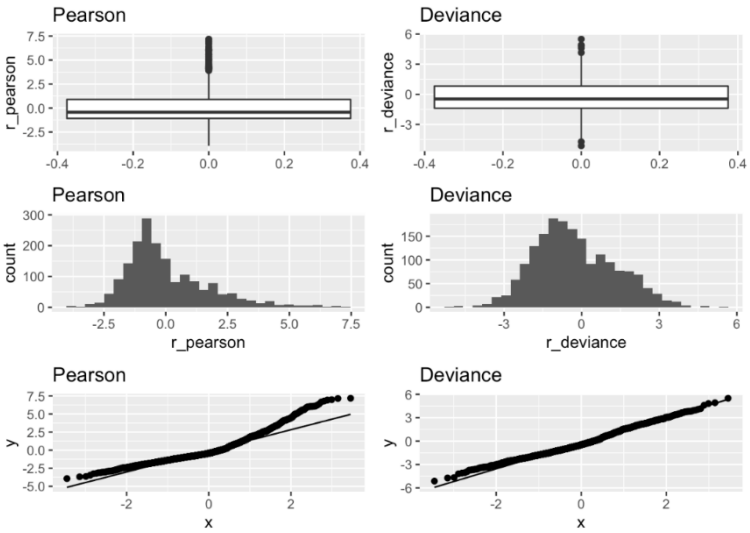
Bonus	Interaction	모형 선택 기준	value	선택된 변수 개수
Catogorical	미포함	AIC	1878541	21
		BIC	1878807	21
	포함	AIC	1863828	189
		BIC	1866225	189
Numeric	미포함	AIC	1879318	16
		BIC	1879521	16
	포함	AIC	1873661	94
		BIC	1874853	94

연간 사고 심도 확률변수가 감마 분포를 따른다고 가정한 후 총 사고건수가 0보다 큰 자료에 대해 Bonus 변수의 연속변수 간주 여부, 교호작용 효과의 포함 여부를 변화시켜가며 모형 적합 후 AIC과 BIC를 기준으로 각각 모형을 선택하였다. AIC를 기준으로 했을 때, bonus가 범주형, 교호작용을 포함한 모형이 선택되었다. 하지만 적합 결과 대부분의 교호작용들의 변수가 유의하지 않게 나왔고, 모형의 복잡도와 모형 해석의 용이함을 고려했을 때 교호작용을 포함하지 않은 1차 선형 모형이 최종 회귀 모형으로 적절하다고 판단하였다. 이때의 AIC값은 1878541이다.



최적 예측 모형을 통해 구한 각 변수 별 효과의 그래프이다. 앞서 구했던 사고 빈도 모델에서는 bonus 감소, kilometres가 증가함에 따라 보험료가 커지는 결과가 나왔으나 해당 모델에서는 그러한 양상을 파악할 수 없다. make가 8일 때 μ값이 높게 측정된다. 이를 통해 make=8은 외제차와 같은 값이 비싼 차종임을 추측할 수 있었다.

2. GLM 모형에서 추정한 평균사고빈도 $\widehat{\lambda_{ijkl}}$, $\widehat{p_{ijkl}}$값과 단순추정값 $\overline{\lambda_{ijkl}}$을 이용한 Pearson 표준화 잔차 및 deviance 잔차를 각 수준별로 구하고 이들 잔차값들의 boxplot; histogram; normal-Q-Q plot을 그려보고 추정모형의 적합도를 검토하시오.



위에서 선택한 최종 회귀 모형을 이용했을 때의 pearson 잔차와 deviance 잔차의 boxplot, histogram, normal q-q plot 그래프이다. Pearson 잔차의 경우 q-q plot에서 오른쪽 끝이 직선에서 벗어나나, deviance 잔차의 경우는 대부분의 점이 직선 위에 있으므로 모형이 잘 적합됐음을 판단할 수 있었다.

3. **평균사고심도에 대한 단순추정값과 GLM을 이용한 추정량의 장단점을 검토하시오.**

총사고건수	총보험금	단순 추정값	GLM 추정값
19	46221	2432.684	4269.806
13	15694	1207.231	4812.187
7	103910	14844.286	4501.011
4	38065	9516.250	4154.741
0	0	NaN	5795.936
0	0	NaN	5663.137
0	0	NaN	5930.034
0	0	NaN	5409.499

총 사고건수가 20보다 작은 경우에 대한 단순 추정값과 glm 모델로 구한 추정값들의 일부를 나타낸 표이다. 총사고건수가 0인 경우 단순 추정값은 추정할 수 없지만 glm 모형의 경우 μ 추정이 가능해 보험료 산출이 가능하다. 또한 총 사고건수가 작을 경우 단순 추정값과 glm 추정값의 차이가 큰데, 이는 샘플 사이즈가 작을수록 단순 추정값은 큰수의 법칙에 의해 실제값과 멀어지기 때문이다. 이를 통해 두 값이 손해율의 측면에서도 차이가 날 것임을 알 수 있다.

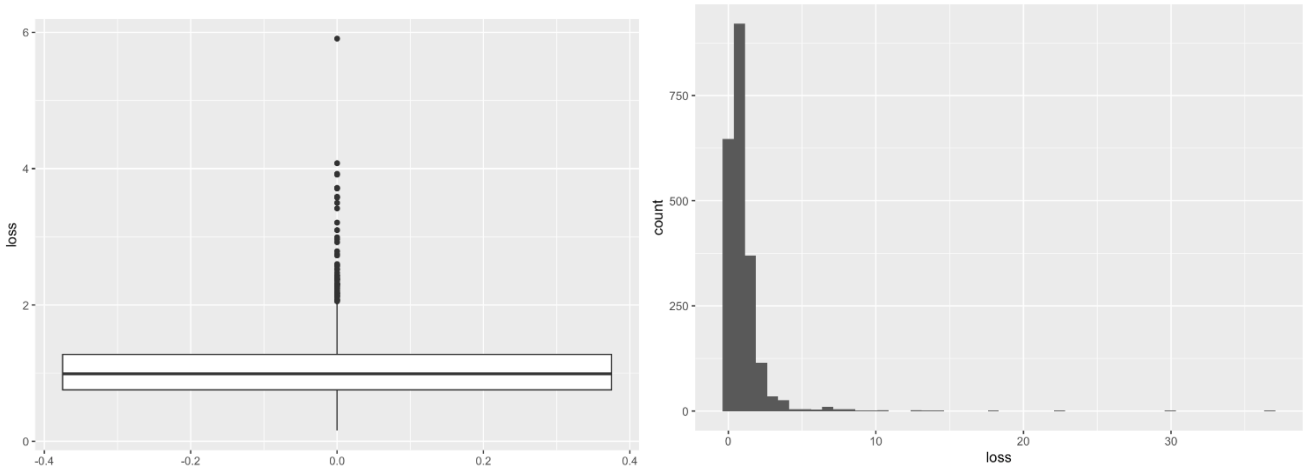
Part 3 : 할증 보험료 자동 계산 Application 개발

1. **위의 두 회귀모형을 이용하여 보험 가입자 및 차량유형 (주행거리, 운전지역, 무사고보너 스, 차종) 별로 차등화 된 연간 자동차 보험료를 산출하는 공식을 제시하시오.**

variable	coefficient		variable	coefficient	
	사고빈도 모델	사고심도 모델		사고빈도 모델	사고심도 모델
Intercept	-1.813	8.395	Bonus4	-0.827	0.057
Kilometres2	0.213	0.025	Bonus5	-0.926	0.034
Kilometres3	0.320	0.021	Bonus6	-0.994	0.070
Kilometres4	0.405	0.043	Bonus7	-1.327	0.116
Kilometres5	0.576	0.039	Make2	0.076	-0.035
Zone2	-0.238	0.023	Make3	-0.247	0.084
Zone3	-0.386	0.048	Make4	-0.654	-0.164
Zone4	-0.582	0.129	Make5	0.155	-0.087
Zone5	-0.326	0.052	Make6	-0.336	-0.039
Zone6	-0.526	0.147	Make7	-0.556	-0.119
Zone7	-0.731	0.023	Make8	-0.044	0.214
Bonus2	-0.479	0.043	Make9	-0.068	-0.055
Bonus3	-0.693	0.069			

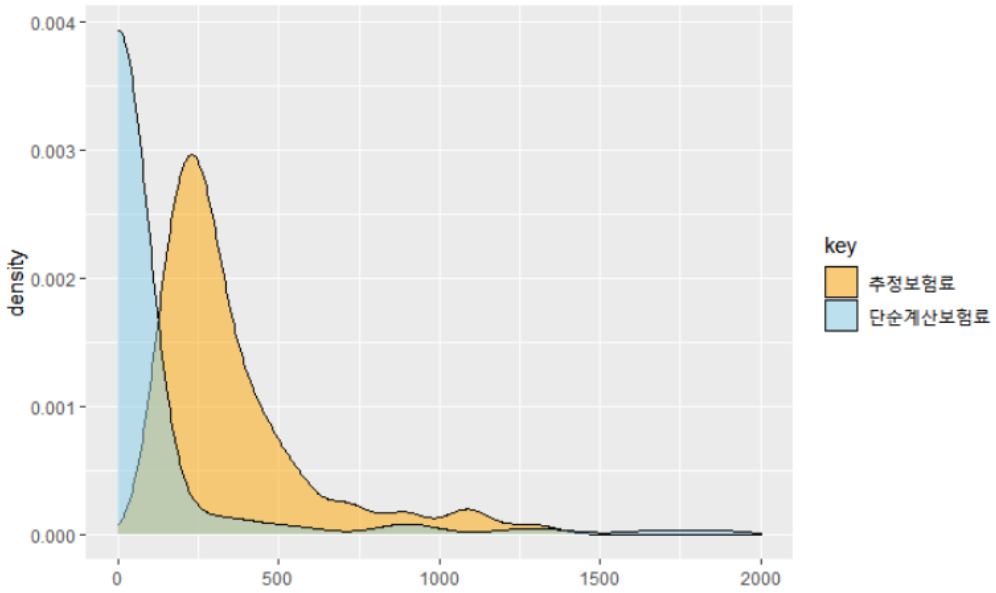
사고빈도 모델과 사고심도 모델의 최적 회귀 모형의 계수이다. 이를 이용해 λ_{ijkl}값과 μ_{ijkl}값을 추정한 후 적정보험료를 계산한다.

2. **GLM을 이용해 계산한 보험료를 적용했을 때 각 수준조합별로 손해율 (실제지급보험료/수입보험료)을 각각 구하고 이들 손해율 값의 boxplot 및 histogram을 그려서 수준별 보험요율의 적절성을 검토하시오.**



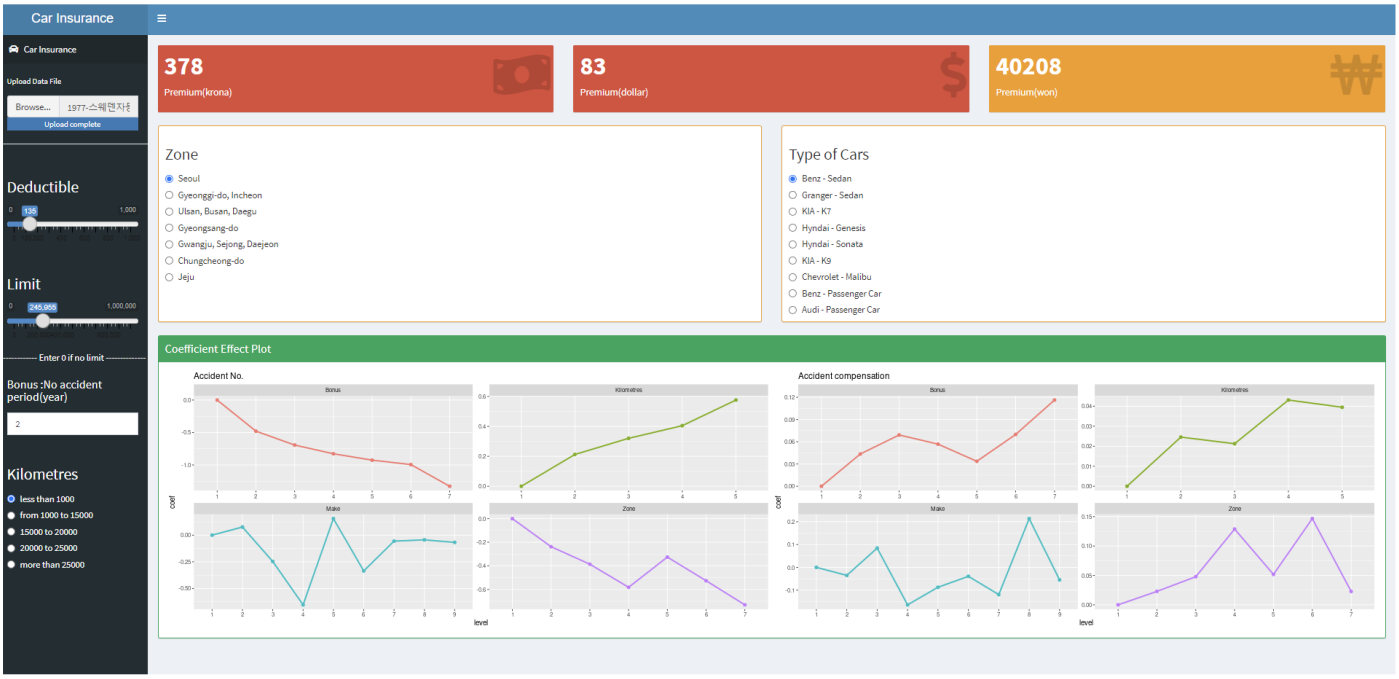
위 공식을 이용하여 보험료를 측정한 후, 실제 지급 보험료와 수입 보험료를 비교한 손해율을 구해 그래프를 그려보았다. 그래프에서 손해율이 대부분 1 근처에 존재함을 확인할 수 있다. 또한 위에서 계산한 보험료를 실제 적용했을 때 전체손해율(총지급보험금/ 총보험료)은 1로 계산되었다. 이를 통해 이상적인 보험료가 산출되었음을 판단할 수 있었다.

3. **각 수준별 단순계산보험료 (지불보험금/보험가입자수)와 GLM을 이용한 추정보험료의 차이점을 설명하고 두 방법의 장단점을 설명하시오.**



보험 가입자가 5명보다 작을 때의 단순계산보험료와 추정보험료의 density 그래프이다. 단순계산보험료의 경우 그래프가 왼쪽으로 치우쳐져 있지만 추정보험료의 경우는 보다 넓게 퍼져있다. 이를 통해 극단값에서 보험료를 계산할 때, 단순추정값을 이용할 경우에는 보험료가 부정확하게 계산되지만 glm 모형을 이용할 경우에는 적절한 보험료를 계산할 수 있음을 알 수 있다.

R-Shiny



GLM 회귀모형을 이용하여 건당 보상한도 및 자기공제액이 있는 경우 적정 보험료를 계산해주는 어플리케이션이다.

(https://2hyeon.shinyapps.io/Car_Insurance/)