Bayesian Statistics for Travel Insurance Modelling

Kevin Hu

April 19, 2021

1 Abstract

This research paper investigates the similarity and differences between Frequentist and Bayesian models in actuarial science setting. We used both methods to predict incidence rates on travel insurance data via logistic regression. We fit four models in total, including two Frequentists and two Bayesians. The first two models used age and trip duration as predictors. The final two models include insurance plan as the third variable. For the Bayesian one, we implemented hierarchical model where plans are encoded in the intercept coefficient.

Overall, both methods yielded similar results in terms of coefficients. When using the fitted models to predict different scenarios, we preferred the Bayesian models as they are less extreme in edge cases. For the models with plan, we preferred the pooled Bayesian model performs than the no-pool Frequentist model as less outliers were predicted.

Unfortunately, our final decision came down to computing power. We would still recommend the Frequentist GLMs since they are much faster to run. Insurance companies deal with large database, and we would not be surprised if a simple Bayesian model would take a few days to run in common cases. With rising computing power, we hope that Bayesian GLMs will be embraced by the actuarial field soon.

2 Introduction

Actuaries have been working with insurance data since the inception of their field. Although many modeling methods have been studied academically over the years, the traditional methods were used until recently. We would sliced our data into different cohorts and investigated them separately to derive trends. Recently, GLMs have gained popularity among the industry. This allows us to systematically analyze all the cohorts without manually perform separate analysis. Given increasing popularity in predictive analytics and big data, more companies are embracing GLMs.

Bayesian statistics are also popular among actuaries as well, but it normally takes a back row seat. Often times, actuaries work with data that are not credible, and often blend experiences that they have with the industrial one. On the other hand, Bayesian GLMs are mostly unheard of in this field. Upon online searches, we almost found no records of Bayesian GLM models in actuarial science setting. This is not surprising as it is not until recently that we have the computing power to construct Bayesian models easily.

3 Data

Finding a public insurance data online is not an easy tasks since they contain confidential information. Fortunately, there was a policy-level travel insurance data provided by Zahier Nasrudin on the Kaggle website[1]. It is a third-party travel insurance database from a Singaporean firm.

3.1 Data Scrubbing

There are 63,326 rows in this dataset, where each row represents a travel insurance policy. The response variate is a binary variable (yes/no) whether the policyholder submitted an insurance claim. Unfortunately, the number of claims per policy and the amount of claims are not available for more analysis. Regardless, this allow us to model incidence rate, which is the number of claimants per policyholders. We can use logistic regression for both Frequentist and Bayesian approaches for this scenario.

There are eleven dependent variables in total, but our interested dependent variables are age, trip duration (in days) and plan. Fortunately, there is no missing data for these fields. We scrubbed the data by subaetting for policies with positive payments only. In other words, policies that were not bought successfully were scrubbed out. We also floored and capped policyholders' ages to 20 and 70 years old respectively for better credibility. Since the study population are not credible at both ends of ages, we do not want them to influence the model. With the same concept, we limit the trip duration to 1 days and 2 years as well. After scrubbing, we ended up with 60,686 rows of data.

3.2 Interested Variables

Our variable of choices are age and trip duration. This is because in real life, travel insurance plans are mostly priced based on age, trip duration and plan. For the simple model, we prefer working with numerical variables first, and introduce the plan categorical variable after. We have attached the sample travel rates below to illustrate the importance of age and duration variables[2].

Premium Table

Premium Table - Travel Guard Silver Plan						
	(worldwide Excluding USA/Canada)			(worldwide Including USA/Canada)		
Age Band Days of Travel	6 Mths-40 Yrs	41-60 Yrs	61-70 Yrs	6 Mths-40 Yrs	41-60 Yrs	61-70 Yrs
1-7	602	644	1014	855	950	1468
8-14	845	918	1369	1216	1315	2049
15-21	955	1060	1759	1275	1528	2633
22-28	1089	1312	2135	1531	1754	3313
29-35	1283	1501	2531	1904	2159	4037
36-47	1522	1849	3198	2332	2790	5209
48-60	1850	2155	4011	3241	4453	6466
61-75	2200	2729	5257	4770	6599	7930
76-90	2557	3226	6708	5703	7096	9820
91-120	3524	4180	9138	5925	7172	15932
121-150	4539	5742	12524	7801	9610	22071
151-180	5568	6868	15292	9832	11851	27417

Figure 1: Sample Travel Insurance Rate Table

Note that there are over 27 travel insurance products in the dataset. We grouped them into four plans below:

- All-Inclusive, which includes medical and trip cancellation benefits.
- Trip-Cancellation, which cover trip costs for unexpected events.
- Rental Vehicle, which cover the cost of vehicle accidents.
- Annual All-Inclusive, which covers All-Inclusive benefits on an annual basis.

[1] https://www.kaggle.com/mhdzahier/travel-insurance
[2] https://www.bookingcounter.com/index.php/welcome/insurance

Observe that the plans above have different risk characteristics, but are not totally independent from each others. We can even think of it in a hierarchical structure, where all policies are issued by the same insurance companies. This motivated us to create the pooled Bayesian model for the plan variable, which we will describe later on.

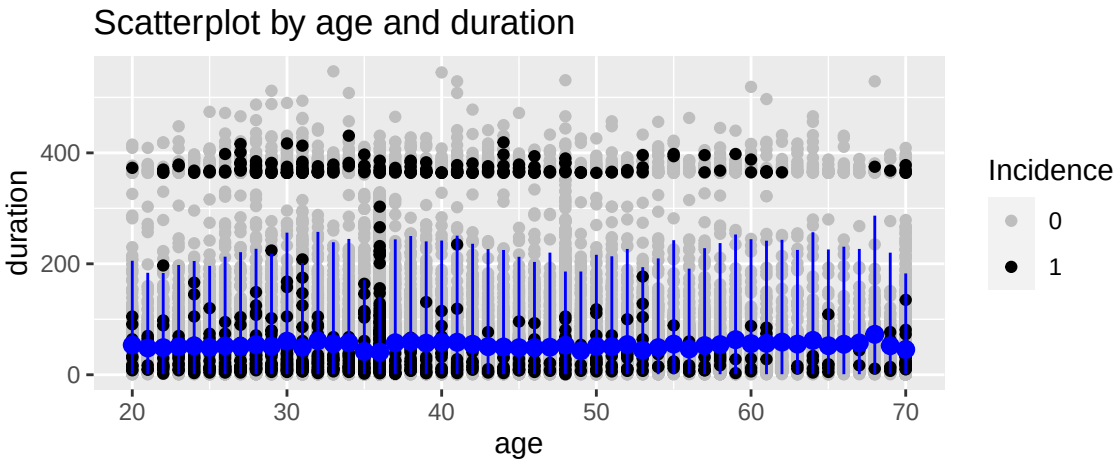
3.3 Variable Characteristics

The summary of data by age and trip duration are shown below. Looks like on average, a traveler is 40 years old with a travel length of 49 days.

Table 1: Summary of numerical variables

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
duration	1	9	22	48	52	547
age	20	35	36	39	43	70

The relationship between age and duration is shown by scatter plot below. Gray dots represent travelers with no claims, and black dot represents travelers with claims. Blue dot and its line represent average trip duration by age and their 95% confidence intervals.



Overall, we do not see an obvious trend between age and travel time. However, it is interesting to see the policies split into two groups of under and over one year of trip lengths. Note that the cohorts above one year are mostly annual all-inclusive travelers.

Recall from Figure 1 that travel insurance rates increase by age and duration. This is a good proxy for us to expect increasing trend in incidence rates by these two variables as well. Note that there are many other external factors affecting the rates such as agent commission structure, but we are still safe to say that higher rates imply higher risk of insurance.

I am comfortably expecting an linear increasing trend in trip duration. However, based on our experience We expect a convex trend by age. There is an inherited risk in younger travelers comparing to the middle ages one. A good proxy to explain this is incidence rates for automobile accidents by age as shown below[3].

[3]https://aaafoundation.org/rates-motor-vehicle-crashes-injuries-deaths-relation-driver-age-united-states-2014-2015/

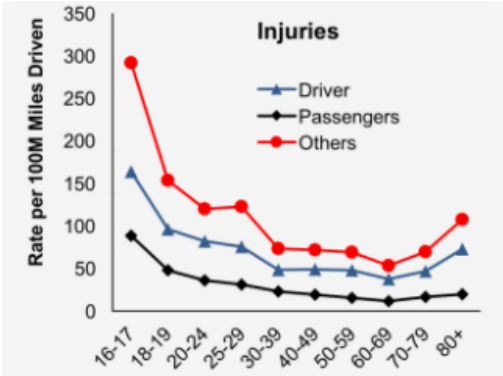
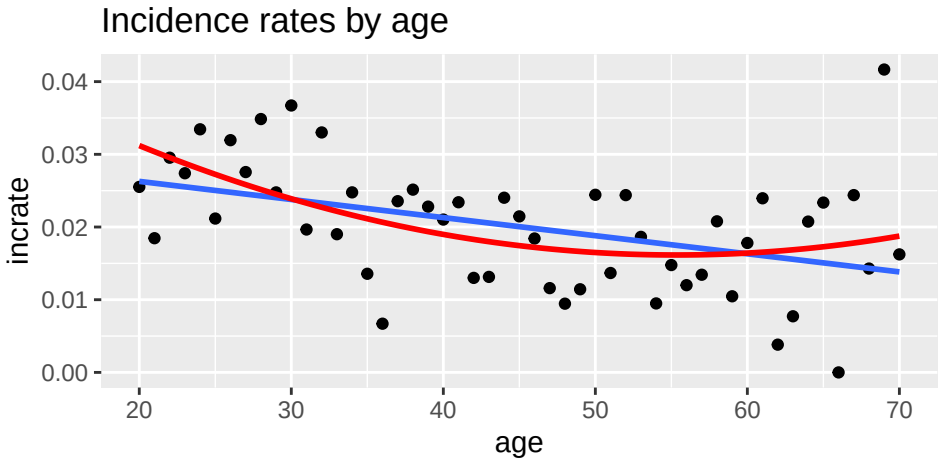
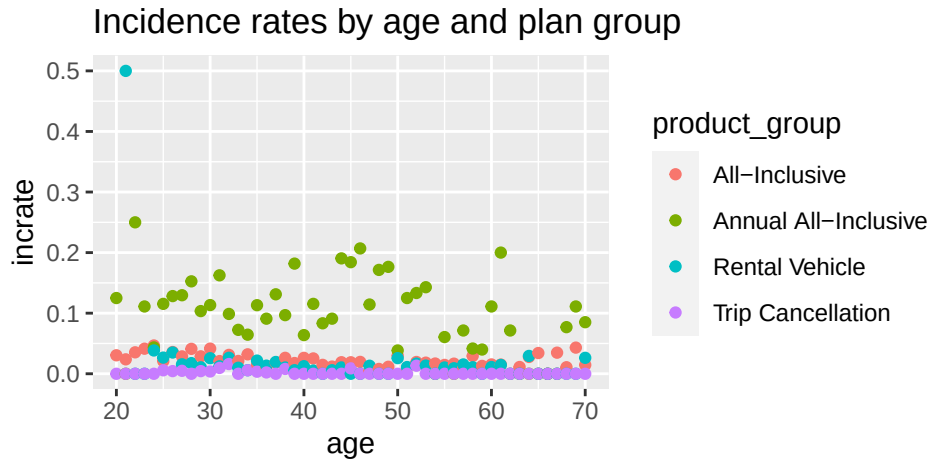


Figure 2: AAA Motor Vehicle Incidence per 100M Miles by Age

We attempt to replicate the figure above by plotting incidence by ages. As shown below, we see a slight convex trend of incidence rates by age. If we were to fit the linear trend (in blue), the coefficient will likely be linear. From our experience, we personally expected a more convex trend line.



Upon investigation below, we found that incidence rates are lowest among the trip cancellation plan, and highest among the Annual All-Inclusive plan. Some plans are volatile because there are not much data, or low credibility in actuarial language. This further encouraged the use of pooling via hierarchical modeling. Normally, it is not possible for actuaries to partially pool experience together in the Frequentist GLM setting, so they would have no choices but to leave all of the plans independent. We will compare the consequences of doing this later on in the paper.



4 Methods

Since our data yields binary responses, we are performing logistic regressions. As mentioned, there will be four models including two Frequentist and two Bayesian GLMs. For each model, y_i represents binary response if a policyholder submit claims (1 for yes and 0 for no). The βs are coefficients that we are trying to fit. We assume that $y_i \sim Bern(p_i)$ where p_i refers to the probability of submitting a claim, which is y_i/n given n as the number of travelers in a cohort. Note that We take log of both age and duration as our Exploratory Data Analysis showed that their distributions become more normally distribute and symmetric.

1. Frequentist simple model

$$logit(p_i) = \beta_0 + \beta_1 \times \log(Age_i) + \beta_2 \times \log(Age_i)^2 + \beta_3 \times \log(Duration_i)$$

2. Bayesian simple model[4]

$$logit(p_i) = \beta_0 + \beta_1 \times \log(Age_i) + \beta_2 \times \log(Age_i)^2 + \beta_3 \times \log(Duration_i)$$

$$\beta_0,\ \beta_1,\ \beta_2,\ \beta_3 \sim N(0, 1)$$

3. Frequentist plan-wise model

$$logit(p_i) = \beta_0 + \beta_1 \times \log(Age_i) + \beta_2 \times \log(Age_i)^2 + \beta_3 \times \log(Duration_i) + \beta_{plan[i]}$$

$$plan[i] \in [4, 5, 6]$$

Where 4 = Annual All-Inclusive, 5 = Rental Vehicle and 6 = Trip Cancellation plans.

4. Bayesian plan-wise model[5]

$$logit(p_i|plan_{j[i]}) = \beta_{0,j[i]} + \beta_1 \times \log(Age_i) + \beta_2 \times \log(Age_i)^2 + \beta_3 \times \log(Duration_i)$$

Where p_i is the probability of policyholder i with plan $j[i]$.

$$\beta_{0j} \sim N(\mu_{\beta_0}, \sigma^2_{\beta_0}), j = 1, ..., J$$

[4]See "res_m1.stan" for the corresponding Stan model
[5]See "res_m2.stan" for the corresponding Stan model

J is the number of plan groups (i.e. All-Inclusive, Rental Vehicle). In other words, we have a pooled intercept based on each travel insurance plan, but they come from the same distribution. The non-pooled coefficients are the same as the base model.

$$\beta_1,\ \beta_2,\ \beta_3 \sim N(0,1)$$

The hyper parameters follow normal distributions below.

$$\mu_{\beta_0} \sim N(0,1),\ \sigma^2_{\beta_0} \sim N^+(0,1)$$

Both Frequentist models above would resemble what actuaries typically use to model incidence assumptions. We will fit them using basic GLM function in R. The simplified Bayesian method and will be fitted using Stan model. For their priors, we used $N(0,1)$ since there are no sources We can refer on. It is common to use normal distribution prior on logistic regression.

The interesting part began in the final model when we introduced hierarchical structure. We recognized that policyholders behave differently according to their plans, but their traveling characteristics are still similar across each others. Thus, we used hierarchical model where the intercepts vary by plan but come from the same distribution. This will be a middle ground between assuming no difference between plans and that all plans are independent across each others.

I will perform several checks on the model. For the Frequentist one, we will analyze the variable significance by p-values. For the Bayesian one, we will provide trace-plots and prior/posterior distribution checks. For both approaches, we will provide heat map of incidence rates for high-level check. Most importantly, we will compare the coefficients between both Frequentist and Bayesian approaches. They will not be identical but should not be too far off.

5 Results

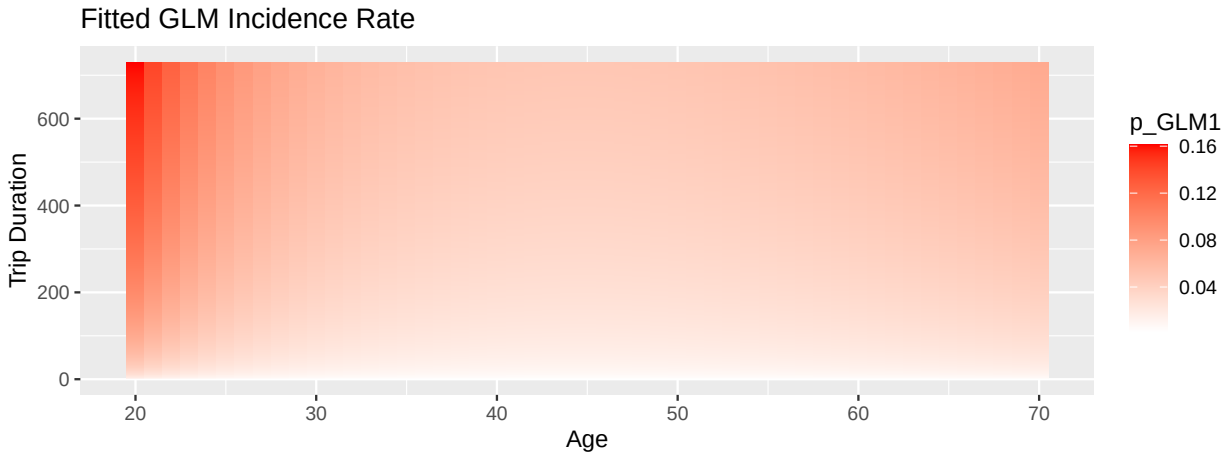
5.1 Frequentist simple model

The coefficients and their p-values are shown below. Overall, we are pretty satisfied with the outputs. The log age coefficients suggested convex incidence trend by age. All variables have low p-values, which signify their importance.

Table 2: Summary of base GLM model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	23.3292476	3.8763714	6.018321	0
logage	-15.4086052	2.1314807	-7.229062	0
I(logage^2)	2.0277293	0.2923416	6.936164	0
logduration	0.4529774	0.0270490	16.746552	0

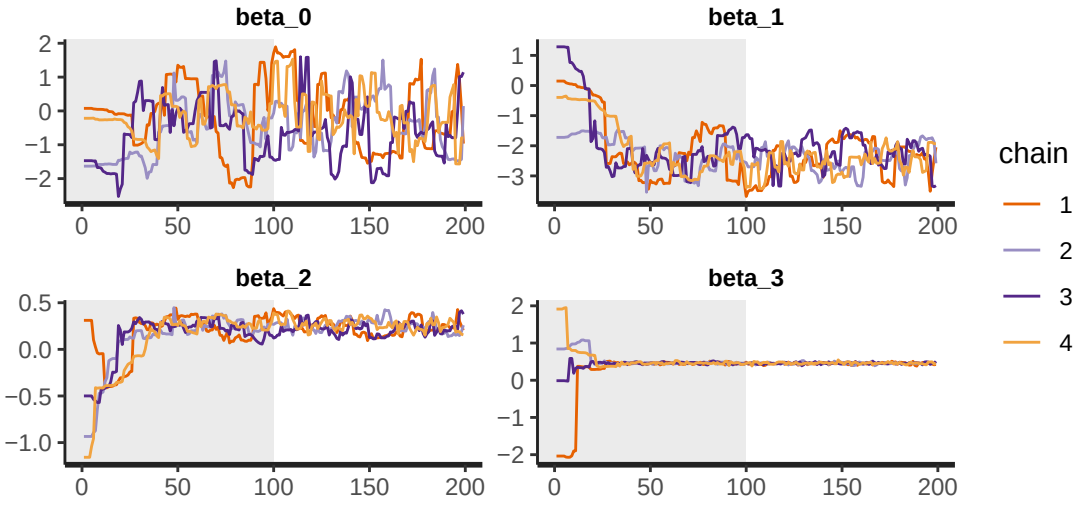
We want to test out the predictive power of our model using the heat map shown below. In the heat map, redder area implies higher incidence rates. Note that we are only plotting heat map from age 20 onward instead of 0. This is because for age and duration at zero, the predicted incidence rate will converge to 100%. Recall that the inverse link function for logistic regression is a sigmoid function, and thus $(1+\exp(-x))^{-1} \to 1$ as $x \to 0$. In real-life we can just assume a flat incidence rates from age 0 to 20, so this is not an issue. Most importantly, we addressed a convex trend of incidence by age, and confirmed that longer duration corresponded to higher incidence.



5.2 Bayesian simple model

We will extend the same logic in our Bayesian model as well. We have written the model in Stan with polynomial age and linear trend coefficients. Please note that we only assign 100 iterations for warm-up and another 100 for the model fitting. Our model is quite large and we have confirmed from the trace plots that the coefficients have already converged early on, so we would like to make the run as efficient as possible.

To be confident with the coefficients, we graphed their trace plots below. Observe that the coefficient converges at 100 iterations. The intercept seem to not be converging as much as we have expected, this might mean that there are still many unexplained variances in the results.



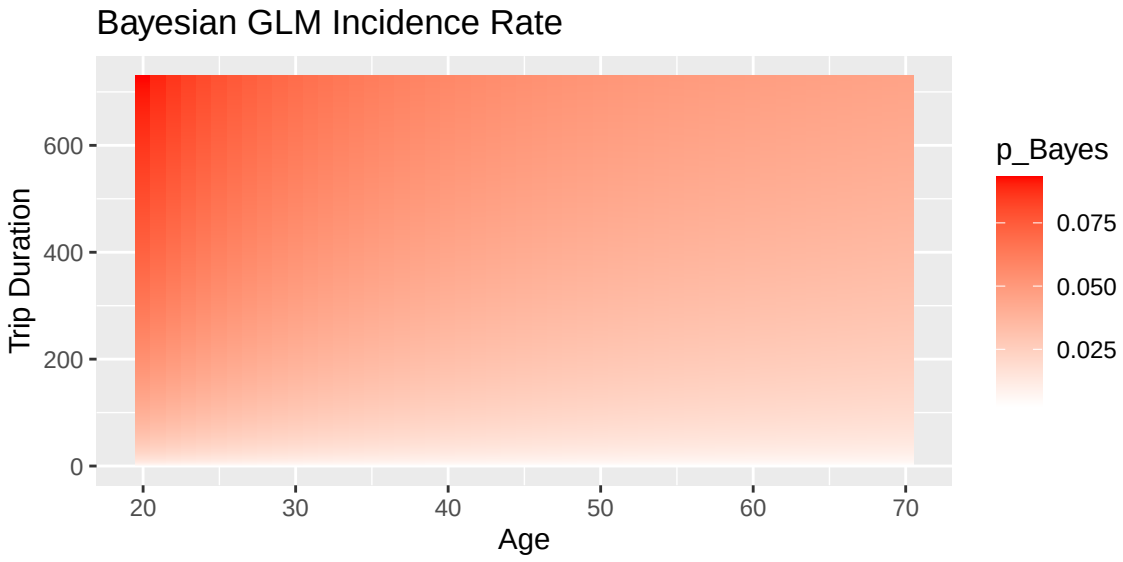
The summary of coefficients are shown below. Although the projected age trends are convex ($\beta_1 < 0$ and $\beta_2 > 0$), its magnitudes are lower than the Frequentist model. For the Frequentist model, both intercept and age coefficients are much larger. This will lead to more extreme results when policyholder ages approach zero. For the Bayesian approach, these coefficients are much smaller. We suspected that it is due to our prior distribution being $N(0, 1)$.

Table 3: Bayesian Regression Coefficient Summary of Statistics

	mean	sd	2.5%	50%	97.5%
beta_0	-0.2823396	0.8771451	-1.8582355	-0.3597662	1.6008901
beta_1	-2.4153945	0.4868134	-3.3879994	-2.3672091	-1.5746233
beta_2	0.2490982	0.0719575	0.1236495	0.2462225	0.3987397
beta_3	0.4561048	0.0261880	0.4015461	0.4561310	0.5065087

The heat map of our Bayesian model is shown below. Unfortunately, the fitted model does not project higher incidence rate as ages increase. Upon further investigation, I found that the model is still convex, but the increasing trend does not come until much later on. This means the incidence rate would only increases when ages become unrealistically large. For our case, it might as well be as good as a decreasing trend.

On the other hand, our data also does not have such a strong convex trend by age to begin with. Looking at the incidence rate by age graph in the Data section, I found that the convexity of the graph was driven by high incidence rate in age 69. It may be that the Bayesian model did not give as much weights to this cohort too. Overall, the incidence rates projected by the Bayesian method is overall lower than the Frequentist model, which I find to be more realistic based on my industrial knowledge. If the incidence rate at any cohort is above 10%, I suspect that the product sold would not be profitable to begin with.



5.3 Frequentist plan-wise model

The summary of coefficients are shown below. Note that the plan coefficients are for Annual All-Inclusive, Rental Vehicle and Trip-Cancellation plans respectively. Observe that all of their p-values are very low, which supports the hypothesis that they are important variables for our model.

Table 4: Summary of plan-wise GLM Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	14.2292729	4.0798992	3.487653	0.0004873
logage	-9.5137513	2.2451846	-4.237403	0.0000226
I(logage^2)	1.1904934	0.3079241	3.866191	0.0001105
logduration	0.1669169	0.0334995	4.982671	0.0000006
as.factor(prodnum)2	1.4233345	0.1255920	11.332998	0.0000000
as.factor(prodnum)3	-0.4086517	0.1162892	-3.514097	0.0004413
as.factor(prodnum)4	-2.0523299	0.1586836	-12.933474	0.0000000

The all-inclusive plan is included in the intercept. The model suggests that annual all-inclusive plan has higher incidence rate since its coefficient is positive. Since the annual plan is technically all-inclusive plan with one year duration, this makes sense to me. On the other hand, the model projects lower incidence rates for rental vehicle and trip cancellation plan. Since these plans are less generous than all-inclusive plan, we find the results reasonable here too.

5.4 Bayesian plan-wise model

Instead of adding another variable for plan group, we use hierarchical model by encoding plans into the intercept. As mentioned, there are two main reasons we are doing this:

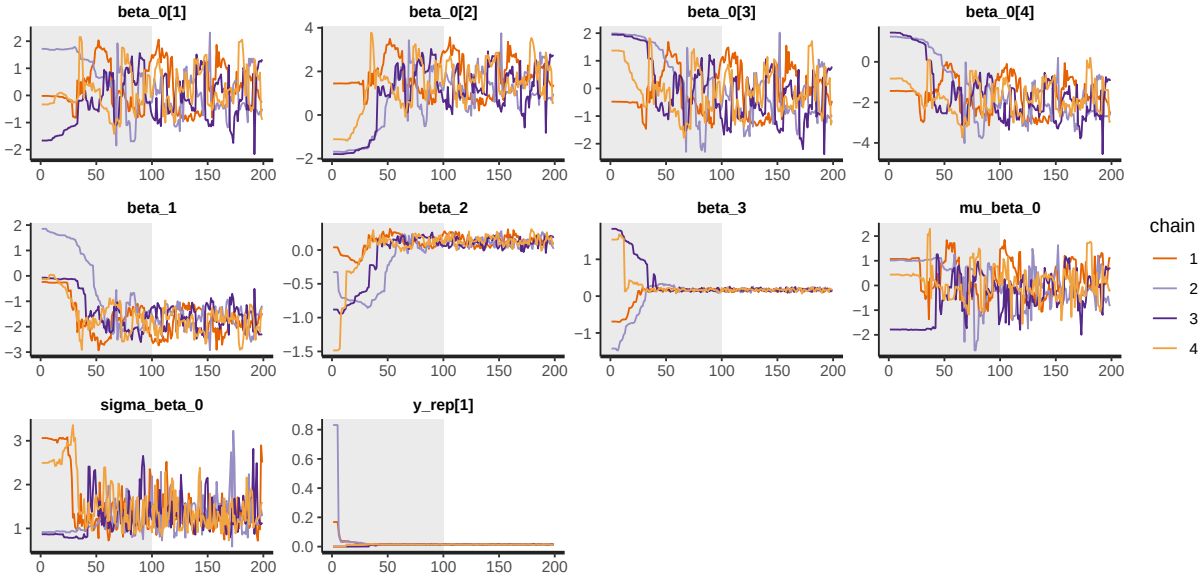
- The data structure is somewhat hierarchical in nature, where travel insurances are from the same insurers.
- Pooling allows information sharing between the model. Since some plans have very low data count, it is more credible if we group them with the larger blocks.

The coefficients are summarized below. Overall, the intercepts and age coefficients are much lower than the Frequentist models as well. This will lead to predicted lower incidence rates as ages approach zero. Note that the intercepts here behave similarly to the plan variables in the Frequentist counterparts. After all, they come from the same variable except that they are not completely independent here.

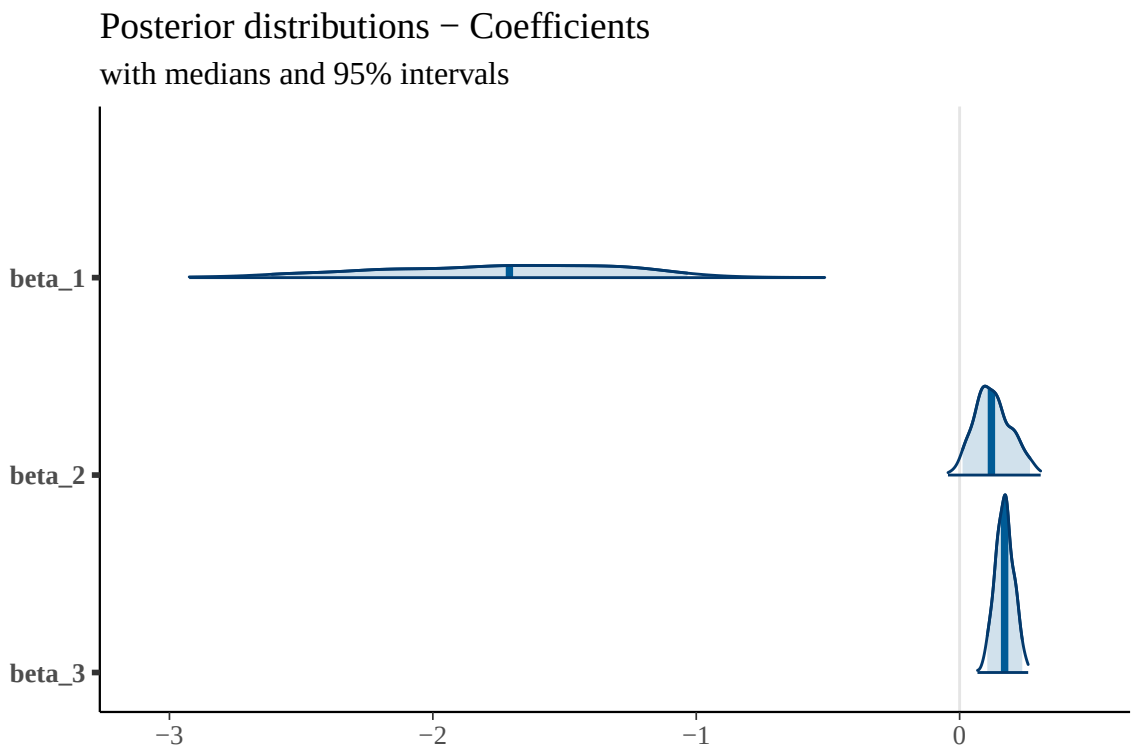
Table 5: Summary of plan-wise Bayesian Model

key	mean	sd	2.5%	97.5%
beta_0[1]	0.1473675	0.8040230	-1.1428231	1.6024999
beta_0[2]	1.5625671	0.8202104	0.1908331	3.0728870
beta_0[3]	-0.2799926	0.8141304	-1.6096380	1.3043925
beta_0[4]	-1.9353114	0.8098821	-3.1778773	-0.3684678
beta_1	-1.7512808	0.4539947	-2.6114642	-1.0337931
beta_2	0.1264147	0.0683190	0.0111190	0.2670884
beta_3	0.1702594	0.0347050	0.1049464	0.2378318
mu_beta_0	-0.0251081	0.7080226	-1.3323321	1.4804354
sigma_beta_0	1.3423321	0.3853141	0.8264085	2.2939463

The trace plots for all estimated parameters are shown below. The top rows contain intercepts for all plans, which are all-inclusive, annual all-iInclusive, rental vehicle and trip-cancellation respectively. The second rows contains coefficients for log age, log age squared and log duration. There are also hyperparameters estimates for intercept mean and standard deviation. Lastly, we used generated quantities for the model so we also predict the incidence rate for each sample.

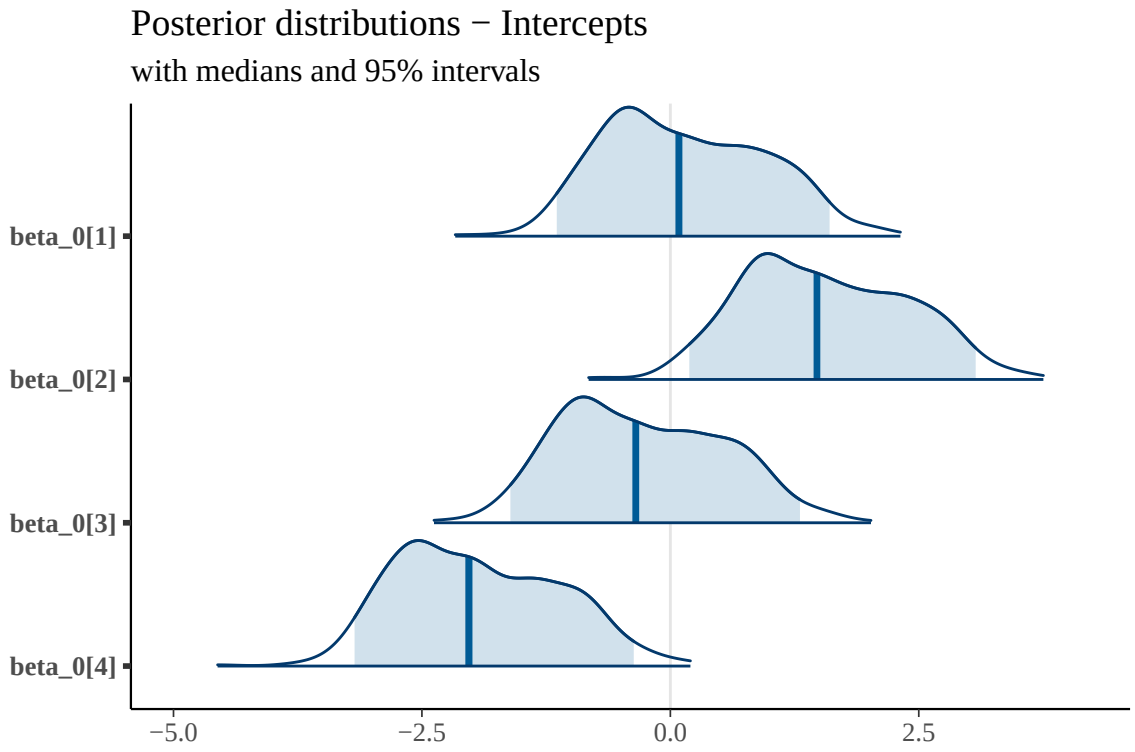


The posterior distributions for the coefficients are shown below. The coefficients for duration (β_3) have very narrow posterior distribution with high probability at median. The coefficients for age (β_1, β_2) have much wider posterior distributions.

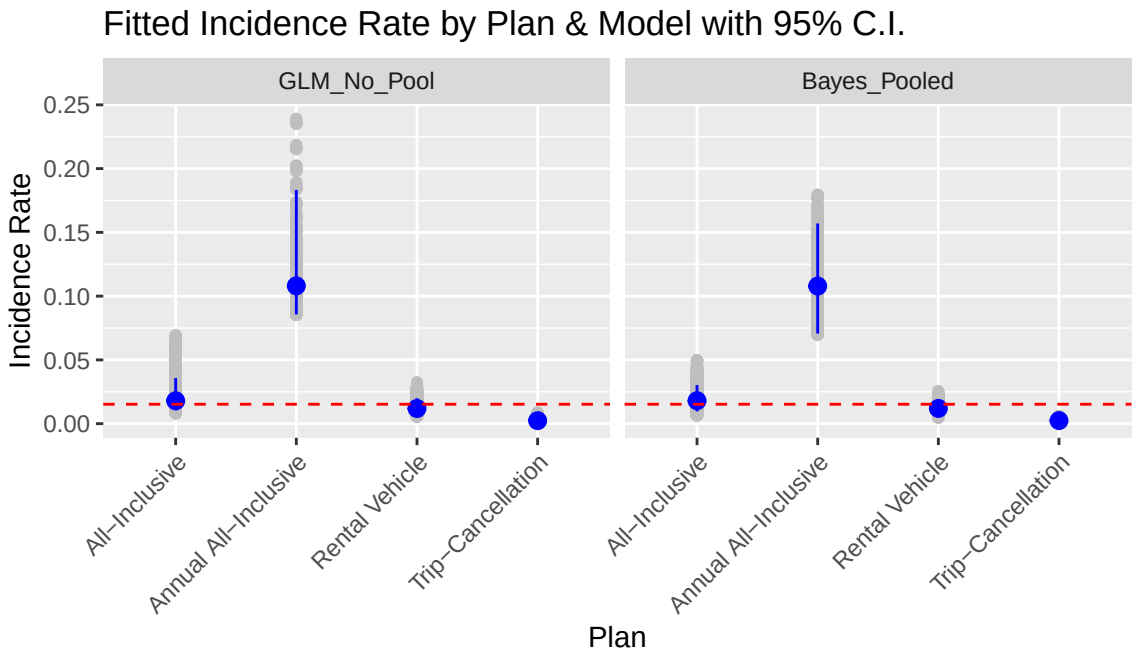


As for the pooled intercepts, we see that Annual All-Inclusive plan ($\beta_{0[2]}$) has higher estimated coefficients than the base All-Inclusive plan ($\beta_{0[1]}$). This is consistent with the Frequentist model where its Annual

All-Inclusive coefficient is positive too. Thus, we are comfortable with the pooled model. The same logic is extended to Rental Vehicle and Trip-Cancellation plans too.



Let's compare the effect of introducing the plan variable between no-pooling method (Frequentist) and pooled method (Bayesian). As we can see below, the pooled model predicted incidence rate closer to the overall mean and are less spread out. There are outliers in the base GLM model with a few fitted incidence rates that are almost as high as 25%. The red dashed line is the overall incidence rates for complete pooling. Overall, we are very satisfied with the pooled results. They do not project overly aggressive incidence rates for the annual all-inclusive plan like the no-pooling model.



The concept above actually is commonly used in actuarial science manually. If we do not want to completely split all experiences independently, we would assign a weighted average of a specific group and its overall cohort. This methodology is called partial credibility. The hierarchical model allows us to automate this process, which saves a lot of modeling efforts.

Lastly, we did a quick check on the sum of fitted incidence rates. There are 924 incidences in the study data. By summing all of the fitted incidence rates, the Frequentist GLM model gives us 924 incidences and Bayesian model gives us 926.5 incidences in total. Overall, both models are performing as expected.

6 Discussion

6.1 Opinions on Results

Overall, we found that the results between Frequentist and Bayesian logistic regressions show some differences. The Frequentist model sucessfully generate convex trend line by policyholder's age, but the projected rates at the edge cases are quite aggressive. On the other hand, Bayesian convex age trend was not within the realistic age range. Then again, the study data itself did not have a strong convex trend by age to begin with. Bayesian models give us more "gentle" results. Looking at the heat-map of projected incidence rates, Bayesians estimates lower incidence rates in more extreme scenarios.

Both models can be extended to include insurance plans. This is where Bayesian model became very useful for actuarial science. By encoding insurance plans in the intercept instead of modeling them independently, we are able to apply credibility weighting automatically. Annual all-inclusive plan benefits from pooling the most. It has the richest benefits, which naturally results in high incidence rates, and also have low policy counts. The Frequentist GLM predicted incidence rates almost has high as 25% for some travelers. Meanwhile, the pooled Bayesian model is more modest in their prediction as it shares information across the plans.

My personal opinion for the actuaries is, unfortunately, to still keep using the Frequentist GLMs. Even though we are a lot more impressed with the Bayesian GLMs, the final decision comes down to computing power. Bayesian models offer more flexibility via hierarchical methodology and prior assumptions, but they take very long time to run. The plan-wise Bayesian model took over three hours to run with just around 60K rows of data. If we were to work with data over 1M rows, we would expect at least one day to finish the run. Frequentist GLMs, on the other hand, took only a few seconds to run. Moreover, results between both Frequentist and Bayesian models convey the same messages, and the potential hierarchical variables (such as destinations) are still not commonly priced in real life. When the computing power catches up with the demand, we can use our coefficients posterior distribution from this paper to help us find the formulate prior knowledge for the next study.

6.2 Potential Future Works

The top priority if given more time will be to refit the age trend for Bayesian model. We still want to see convex trend across reasonable age ranges. Some potential ideas are assigning priors with higher ranges (i.e. $N(0, 5)$) so the model has more chances experimenting with more coefficient ranges. Otherwise, splines can also be implemented as well.

There are also other variables worth exploring. For example, the destination variables would be a perfect candidate for hierarchical modeling. However, since most insurers do not price travel insurance by travel destinations, there is less motivations to pick it for the this paper. Gender is also another variable that is worth experimenting. Unfortunately, almost as much as 80% of the policies have missing gender field.

Another data field that we wish we have is the claim amount per claimant. This will allow us to fit model for claims severity too. In actuarial science, once we have both incidence rates and claims severity, we can multiply them to calculate expected losses per policyholder. It will be a crucial building block for pricing travel insurances.