STSB6816 Test 2 of 2024

Mathematical Statistics and Actuarial Science; University of the Free State

2024/05/23

Time: 170 minutes; Marks: 45

MEMORANDUM

Instructions

- Answer all questions in a single R Markdown document. Please knit to PDF or Word at the end and submit both the PDF/Word document and the ".Rmd" file for assessment, in that order.
- Label questions clearly, as it is done on this question paper.
- All results accurate to about 3 decimal places.
- Show all derivations, formulas, code, sources, and reasoning.
- Intervals should cover 95% probability unless stated otherwise.
- No communication software, devices, or communication capable websites may be accessed prior to submission. You may not (nor even appear to) attempt to communicate or pass information to another student.
- Use of AI tools must be disclosed and summarised.

Introduction

The data is provided at https://ufs.blackboard.com. It is a manipulated and distorted sample of events for a group life insurer, that is an insurer who sells insurance to a company to cover the lives of all their employees as a group. Every employee in a company pays the same premium and then when an individual dies their beneficiaries and the company get a small payout to help cover the resulting expenses.

The sample was distorted by removing about 200000 non-events randomly. This should not affect the statistical significance of independent variables, but does massively inflate the rate of deaths. Pretend that this insurance is for mercenary companies in a war zone, then the numbers will make sense.

Your objective is to fit an appropriate generalised linear mixed effects model on the target events. The choice of explanatory variables is largely up to you, but must include at least age last birthday (ALB) and gender as fixed effects, along with clients as random intercepts. Use of log salary is optional, as is interactions, but do not use any of the other variables.

Question 1

1.1) You should fit two of the following models and compare them to determine which best fits the data:

1. Poisson regression (log link)
2. Logistic regression (logit link)
3. Negative Binomial (Type II) regression (log link)

Additionally, you must use at least two selections of explanatory variables and compare those. So, in total, you must fit at least 3 models and compare them. Explain each model structure used briefly and explain what the result of the model comparison implies. **[20]**

[Hint: if you have trouble comparing models for any reason then use any one reasonable fit to answer the rest of the test, as if it was the best model.]

```
library(tidyverse)
library(rstan)
library(rstanarm)
options(mc.cores = 3)

"STSB6816Test2Data2024.xlsx" |> openxlsx::read.xlsx("TestData") -> d

fits <- list(
  poisson_fit1 = stan_glmer(Target ~ ALB*Gender + log_salary + (1|Client),
                    data = d, family = poisson(link = "log"), iter = 6000),
  poisson_fit2 = stan_glmer(Target ~ ALB + Gender + (1|Client),
                    data = d, family = poisson(link = "log"), iter = 6000),
  logit_fit1 = stan_glmer(Target ~ ALB*Gender + log_salary + (1|Client),
                    data = d, family = binomial(link = "logit"), iter = 6000),
  logit_fit2 = stan_glmer(Target ~ ALB + Gender + (1|Client),
                    data = d, family = binomial(link = "logit"), iter = 6000),
  nb_fit1 = stan_glmer.nb(Target ~ ALB*Gender + log_salary + (1|Client),
                    data = d, iter = 6000),
  nb_fit2 = stan_glmer.nb(Target ~ ALB + Gender + (1|Client),
                    data = d, iter = 6000)
)

library(loo)
fits |> lapply(\(fit) {loo(fit, cores = 3)}) |>
  loo_compare() -> comparison
model_order <- order(rownames(comparison))
rownames(comparison) <- names(fits)[model_order]
comparison |> knitr::kable(digits = 2)
```

	elpd_diff	se_diff	elpd_loo	se_elpd_loo	p_loo	se_p_loo	looic	se_looic
poisson_fit1	0.00	0.00	-106.48	16.35	10.25	2.08	212.96	32.71
poisson_fit2	-0.04	0.94	-106.52	16.40	8.83	1.74	213.05	32.81
nb_fit1	-0.84	0.54	-107.32	16.62	9.19	1.87	214.64	33.23
nb_fit2	-1.07	1.05	-107.55	16.71	8.16	1.62	215.11	33.42

	elpd_diff	se_diff	elpd_loo	se_elpd_loo	p_loo	se_p_loo	looic	se_looic
logit_fit1	-1.77	0.93	-108.26	16.87	8.74	1.80	216.51	33.73
logit_fit2	-1.85	1.24	-108.34	16.92	7.62	1.52	216.67	33.85

Fitting a model that meets the criteria [6]. Fitting a second model that has a different distribution [4]. Fitting a third model that has a different set of explanatory variables [4]. Comparing the models objectively [6]. The Poisson model and the negative binomial model use log links, which are in line with the expected relationship between age and mortality, but the negative binomial allows for more dispersion. The binomial model uses a logit link which has an S shape so may not extrapolate as well to higher ages. The models are essentially equivalent, within uncertainty. The model listed first can be seen as the best but no definitive statement in this regard should be made. Any valid model may be used going forward.

For the best model:

1.2) Analyse the fixed effect parameter summaries in depth. **[5]**

```
summary_stats <- function(sims_vector, width = 0.95) {
  v <- sims_vector
  sym_int <- quantile(v, c((1-width)/2, (1+width)/2))
  dens <- density(v)
  c(
    Mean = mean(v),
    Median = median(v),
    Mode = dens$x[which.max(dens$y)],
    L = sym_int[1],
    U = sym_int[2]
  )
}

post_sims <- rstanarm::as_draws(fits$poisson_fit1)
vars_of_interest <- seq_len(ncol(post_sims) - 3)
summary_table <- vars_of_interest |> sapply(\(j) {
  post_sims[[j]] |> summary_stats()
})
summary_table <- summary_table |> round(3) |> t()
data.frame(Variable = names(post_sims)[vars_of_interest], summary_table) |>
  knitr::kable(digits = 3)
```

Variable	Mean	Median	Mode	L.2.5.	U.97.5.
(Intercept)	-2.710	-2.678	-2.315	-6.580	0.996
ALB	0.014	0.014	0.014	-0.042	0.069
GenderM	-4.459	-4.413	-4.593	-8.981	-0.287
log_salary	-0.100	-0.104	-0.146	-0.449	0.264
ALB:GenderM	0.067	0.067	0.075	-0.023	0.159
b[(Intercept) Client:A]	0.418	0.365	0.057	-0.783	1.756
b[(Intercept) Client:B]	-0.115	-0.055	-0.013	-1.608	1.108
b[(Intercept) Client:F]	0.324	0.291	0.050	-0.648	1.384
b[(Intercept) Client:G]	-0.509	-0.437	-0.071	-1.832	0.467

Variable	Mean	Median	Mode	L.2.5.	U.97.5.
b[(Intercept) Client:J]	0.846	0.815	0.862	-0.207	2.176
b[(Intercept) Client:N]	-0.659	-0.549	-0.093	-2.285	0.365
b[(Intercept) Client:O]	-0.828	-0.746	-0.605	-2.367	0.206
b[(Intercept) Client:P]	0.165	0.132	0.020	-0.983	1.350
b[(Intercept) Client:Q]	-0.068	-0.045	-0.010	-1.079	0.878
Sigma[Client:(Intercept),(Intercept)]	0.802	0.572	0.323	0.008	2.997

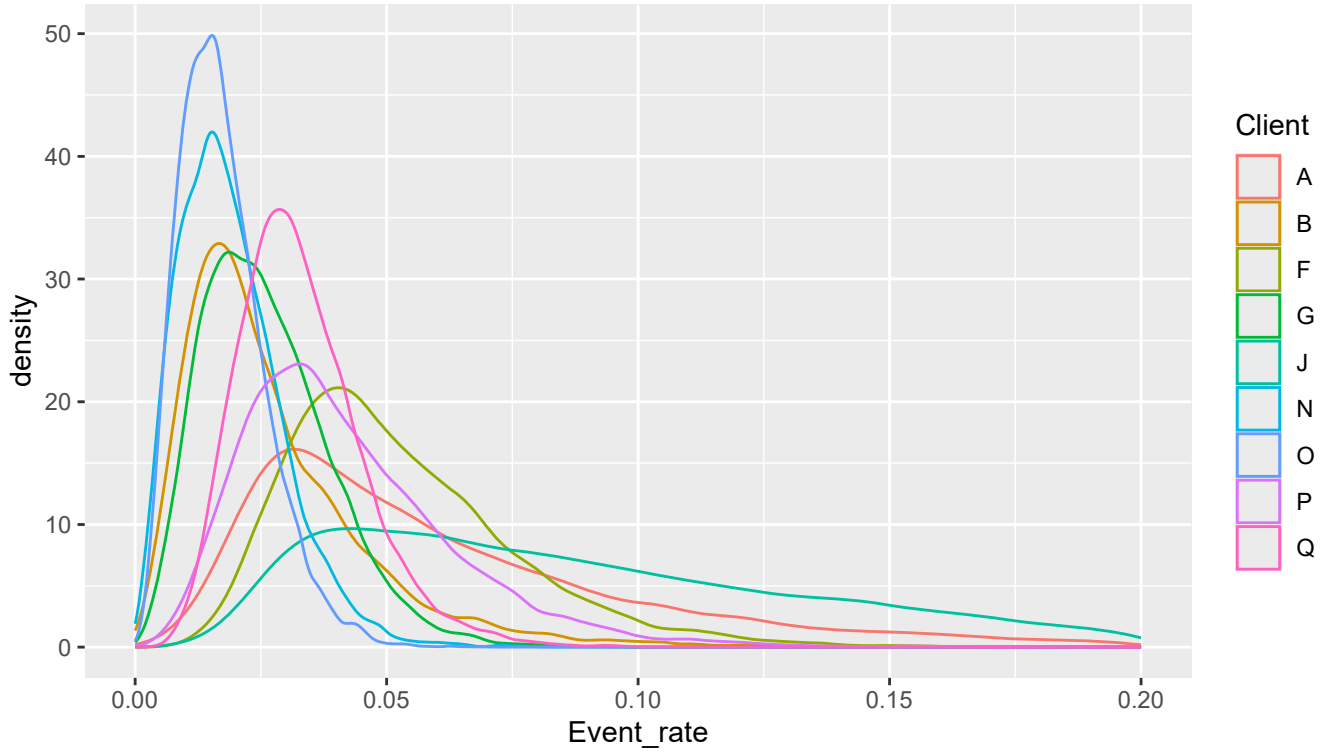
The only variable that typically comes out as significant is gender, with males having lower expected mortality. This may be a false positive or due to some selection effect [5]. The model suggests a positive relationship with age, which makes sense, so it is surprising that it is not more significant.

1.3) Compare the distributions of the predicted event (death) rates of the different clients visually (adjusted for client size). State which client has the highest expected event rate (on any basis). **[7]**

[Hint: The key steps are: obtaining an explanatory matrix, multiplying by the parameter matrix, and applying the appropriate transformation. You only need to do it for Client N in order to answer the questions to come.]

```
Xmatrix <- cbind(model.matrix(~ALB*Gender + log_salary, data = d),
                 model.matrix(~Client -1, data = d))
clients <- d$Client |> unique()
event_rates <- clients |> sapply(\(client) {
  client_rows <- which(d$Client %in% client)
  colMeans(exp(Xmatrix[client_rows, ] %*% t(post_sims[seq_len(ncol(Xmatrix))])))
}) |> as.data.frame() |>
  pivot_longer(everything(), names_to = "Client", values_to = "Event_rate")

event_rates |> ggplot(aes(x = Event_rate, colour =  Client)) +
  geom_density() + xlim(0, 0.2)
```



```
event_rates |> group_by(Client) |>
  summarise(Expected_Value = mean(Event_rate)) |>
  arrange(desc(Expected_Value)) |> knitr::kable(digits = 3)
```

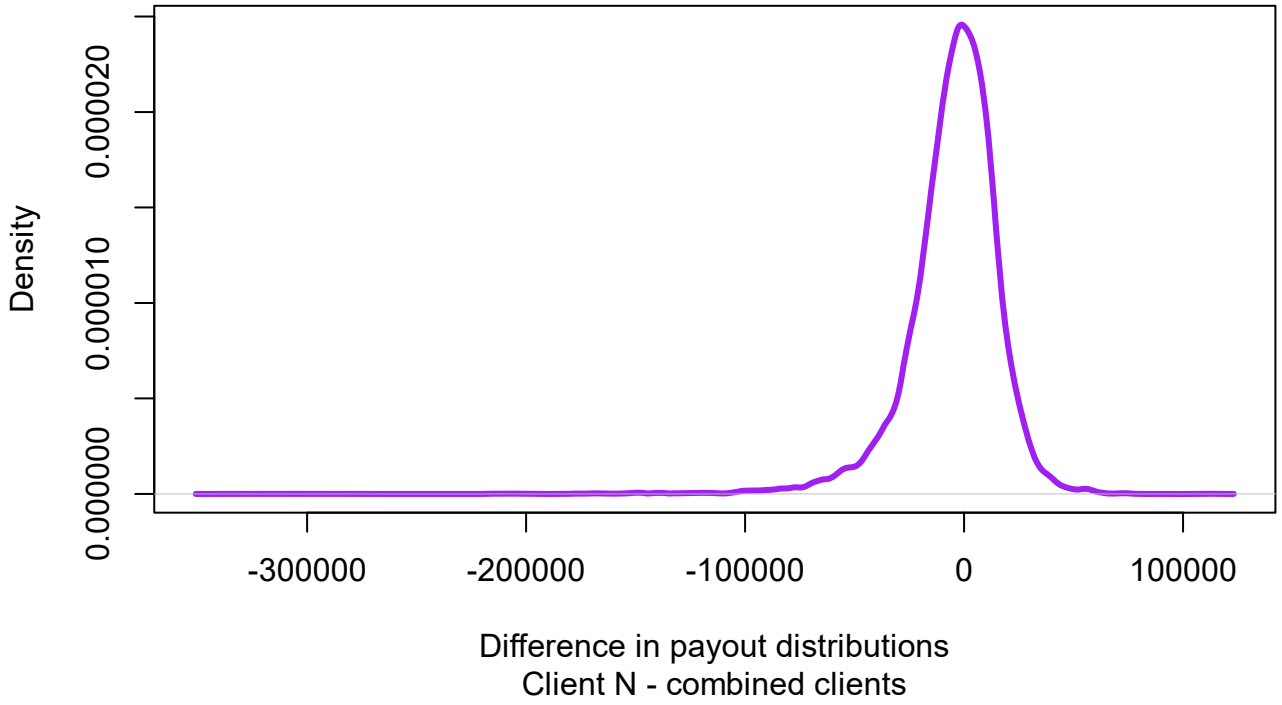
Client	Expected_Value
J	0.099
A	0.064
F	0.053
P	0.042
Q	0.033
B	0.027
G	0.026
N	0.019
O	0.017

First, isolate the data of a single client [1]. Multiply the data by the coefficients and exponentiate [2]. Average over the individuals within client [1]. Repeat for multiple clients [2]. Using the plot, or a table such as above, Client J has the longest tail and thus the highest expected value [1].

1.4) Compare the distribution of the standardised predicted payouts for Client N versus the average client (per individual) as follows: subtract the distributions and illustrate the resulting density. **[7]**

[Hint: the standardised predicted payout for a client is the dot product of the event probabilities and the sum assured values, divided by the number of individuals within the client.]

```
Xall <- model.matrix(~ALB*Gender + log_salary, data = d)
Yall <- exp(as.matrix(post_sims[seq_len(ncol(Xall))]) %*% t(Xall))
payout_all <- Yall %*% d$Final_SA / nrow(d)
client_rows <- which(d$Client %in% "N")
Y1 <- exp(as.matrix(post_sims[seq_len(ncol(Xall))]) %*% t(Xall[client_rows, ])
          + post_sims$`b[(Intercept) Client:N]`)
payout_1 <- Y1 %*% d$Final_SA[client_rows] / length(client_rows)
payout_diff <- payout_1 - payout_all
par(mar = c(6,5,1,1))
payout_diff |> density() |> plot(main='', xlab = "Difference in payout
distributions",
                                 sub = "Client N - combined clients",
                                 col = 'purple', lwd = 3)
```



Predicted payouts for all clients, no RE [2]. Predicted payouts for N (using RE) [1]. Using the full uncertainty of the parameter values [2]. Differencing and presenting a density plot [2].

1.5) Assume that premiums are set for each client at the 60% quantile of the expected payout for that client, taking the fixed effects into account but not the random effects. Calculate the premium for Client N, standardised by client size. **[3]**

```
Y_N <- exp(as.matrix(post_sims[seq_len(ncol(Xall))]) %*% t(Xall[client_rows, ]))
payout_N <- Y_N %*% d$Final_SA[client_rows] / length(client_rows)
cat('\nThe premium for Client N is', round(quantile(payout_N, 0.6), 0), 'per
individual.\n')
```

```
|
| The premium for Client N is 48246 per individual.
```

Calculating expected payout distribution without client effect [2]. Reporting 60% quantile [1].

1.6) Consider again Client N. If they were a new random client instead of an existing client, how much higher would you set their premium? Repeat the premium calculation including random draws from the between client distribution and produce an exact figure per individual. **[3]**

```
Y_N_pred <- exp(as.matrix(post_sims[seq_len(ncol(Xall))]) %*% t(Xall[client_rows, ])
+
             rnorm(nrow(post_sims), 0,
post_sims$`Sigma[Client:(Intercept),(Intercept)]`))
payout_N_pred <- Y_N_pred %*% d$Final_SA[client_rows] / length(client_rows)
cat('\nThe higher premium for Client N is', round(quantile(payout_N_pred, 0.6), 0),
'per individual.\nThus the increase is', round(quantile(payout_N_pred, 0.6) -
quantile(payout_N, 0.6),0), 'per individual.\n')
```

```
|
| The higher premium for Client N is 48735 per individual.
| Thus the increase is 489 per individual.
```

Calculating expected payout distribution with random client effect [2]. Reporting 60% quantile [1].

Points total

The points on the test add up to **45**
