Bayes Data Analysis classification

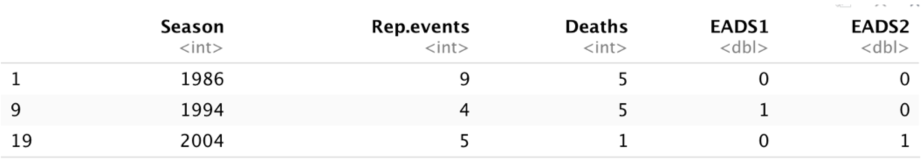
Problem 01-

01-(a): Exploratory on avalanche data

Exploratory analysis: Define a ratio = deaths/rep.events has a decreasing trend over the periods considered. During 1986-1993 and 1994-2003, when reported events increased suddenly, ratio would rush up, even exceed over 1 in 1991 and 1994, therefore this ratio fluctuated a lot during 1986-1993 and 1994-2003. But during 2004-2019, the ratio decreases to a low level (around 0.4) but still has little bumps.

1991 has the largest deaths and corresponding ratio, 2005 has the largest reported events but has the lowest ratio.

Ratio line plot is given in Figure 1. Line plot of reported events and deaths is given in Figure 2.



	Season <int>	Rep.events <int>	Deaths <int>	EADS1 <dbl>	EADS2 <dbl>
1	1986	9	5	0	0
9	1994	4	5	1	0
19	2004	5	1	0	1

Table 1: dataset rows relative to the years 1986, 1994, 2004.

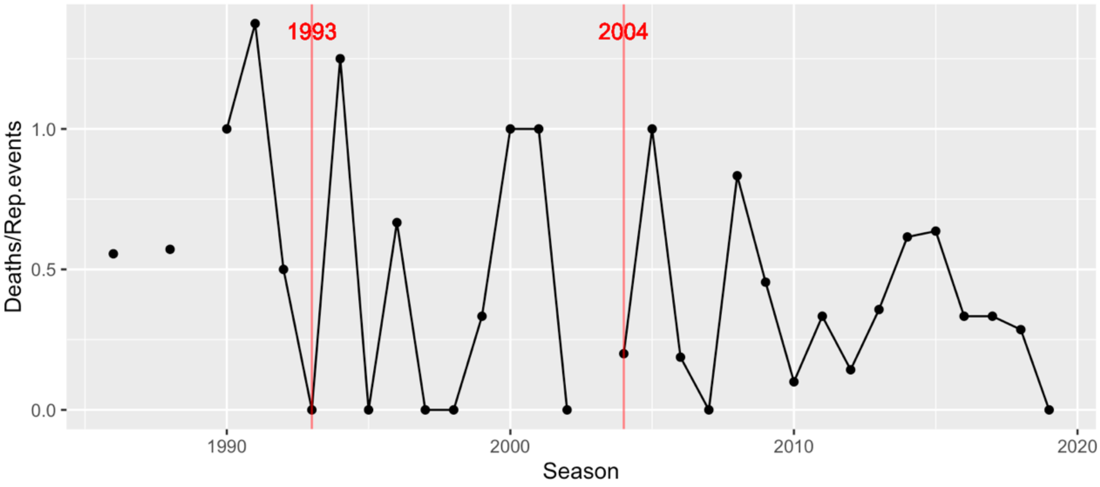


Figure 1: Line plot of death ratio (deaths/reported events)

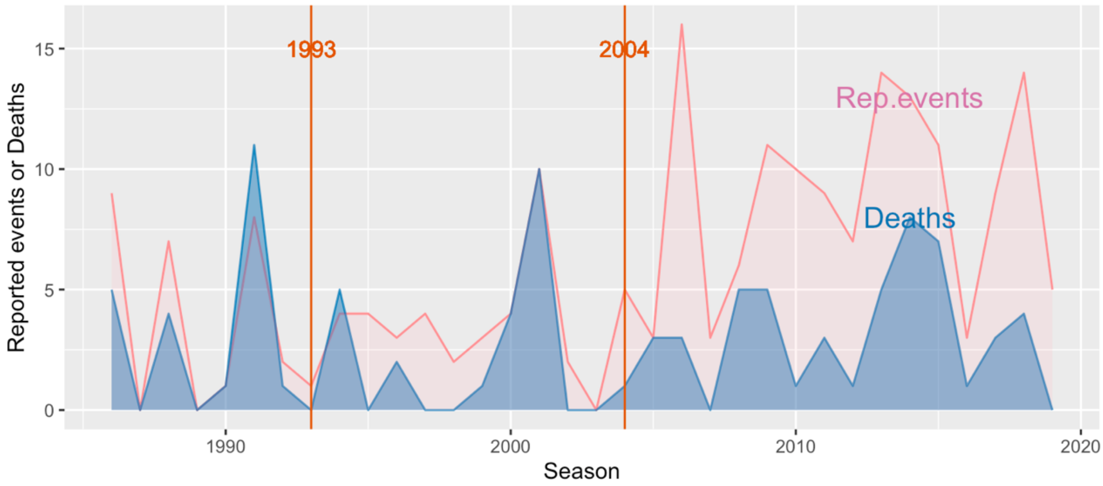


Figure 2: Temporal evolution of the number of avalanches and deaths

01-(b): fitting a GRM on avalanches events

Parameter interpretation: Consider the analysis of the relationship between the number of Deaths (response variable) and the covariates “Reported events” and two dummies “EADS1”, “EADS2”. I set the following generalized linear model with a logarithm link connecting the response mean with the linear combination of the covariates:

$$\log(\mu_i) = \beta_0 + \beta_1 \widetilde{Rep.event_i} + \beta_2 EADS1_i + \beta_3 EADS2_i \qquad (1)$$

$$\mu_i = \exp\{\beta_0 + \beta_1 \widetilde{Rep.event_i} + \beta_2 EADS1_i + \beta_3 EADS2_i\}$$

where $\widetilde{Rep.event_i}$ represents the centered reported events. The second line of Equation (1) highlights the multiplicative effect of the covariates on the mean μ_i.

After running JAGS, Brooks-Gelman-Rubin statistic of β are all around 1 and there’s no trend in trace plots, no severe lag in ACF plots, which indicates MCMC converged.

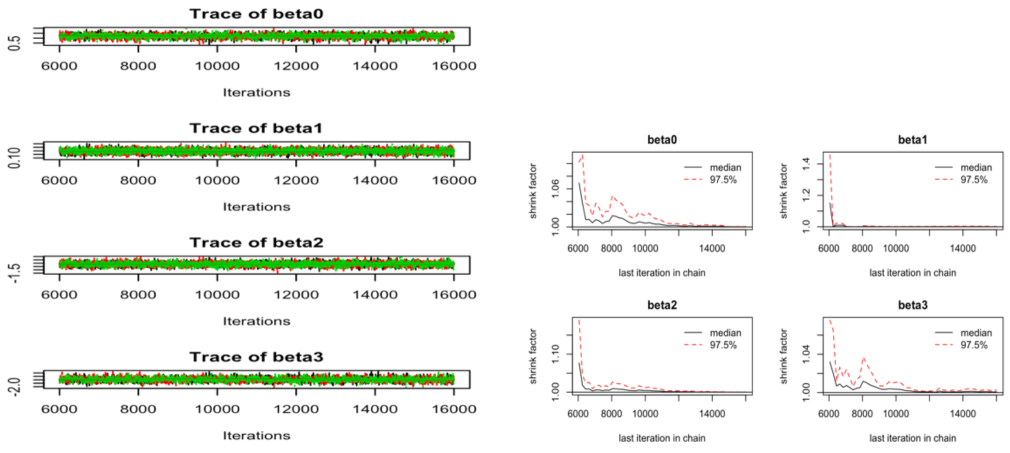


Figure 3: Trace plots and Gelman plots for parameters in model 01-(b)

Interpretations of results are as follow:

- $E[\exp\{\beta_0\}]$=3.45 is the expected number of deaths for a year between 1986 -1993.
- A year with a reported events one unit more than the mean has an expected number of deaths increased of $E[\exp\{\beta_1\}]$ =1.22, while the other variables are hold constant.
- A year between 1994-2003 has an expected number of deaths increased of $E[\exp\{\beta_2\}]$ =0.91, while the other variables are hold constant.
- A year after 2004 has an expected number of deaths increased of $E[\exp\{\beta_3\}] =$ 0.41, while the other variables are hold constant.

01-(c): Posterior probabilities and mean

i. The probability of observing μ<15 when reported events = 20, year = 2020 (thus EADS1=0, EADS2 = 1) is 0.188.

$$\mu_i = \exp\{\beta_0 + \beta_1 \times (20 - mean(rep.events) + \beta_2 \times 0 + \beta_3 \times 1\}$$

ii. Posterior mean of μ >1 in period1 (EADS1=0, EADS2=0) is 0.82; in period2 (EADS1=1, EADS2 = 0) is 0.67; in period3 (EADS1=0, EADS2 =1) is 0.018. Probabilities shrink through these periods.

01-(d): Check proposed prior feasibility

Calculation:

- Based on $\phi \sim Lognormal(\mu_0, \sigma_0^2)$, therefore:

$$p(x) = \frac{1}{\beta\exp(x-\mu_x)\sigma_0\sqrt{2\pi}}\exp\left(-\frac{(\ln(\beta\exp(x-\mu_x)-\mu_0))^2}{2{\sigma_0}^2}\right) \times |\beta\exp(x-\mu_x)'| \qquad (2)$$

Where x is Reported events, $\mu_0 = 0$, $\sigma_0 = 2$, $\mu_x = mean(Rep.events) = 5.97$, $\beta = \beta_{rep.events}$.

(2) can be simplified into:

$$\frac{1}{\sigma_0\sqrt{2\pi}}\exp\left(-\frac{(x+\ln(\beta)-\mu_x-\mu)^2}{2{\sigma_0}^2}\right)$$

that is to say $X \sim N(\mu_x + \mu_0 - \ln(\beta), \sigma_0^2)$.

- Similarly, I derive pdf of β :

$$\frac{1}{\beta\sigma_0\sqrt{2\pi}}\exp\left(-\frac{(\ln(\beta)+x-\mu_x-\mu)^2}{2{\sigma_0}^2}\right)$$

That is to say $\beta_{rep.events} \sim Lognormal(\mu_x + \mu_0 - x, \sigma_0^2)$.

- According to experts, x is usually between 5 and 15, I assume this means expectation of x is 10 and standard deviation is 5; $\beta_{rep.events}$ is between $\frac{1}{4}$ and 4, I assume this means expectation of $\beta_{rep.events}$ is 2.1 and standard deviation is 1.87. My computation base on $\phi \sim Lognormal(\mu_0, \sigma_0^2)$ indicates that expectation of x is $\mu_x + \mu_0 - \ln(\beta) = 5.97 - \ln(\beta_{rep.events})$, which means if lognormal prior for ϕ is appropriate, expert' s expectation of x (which is 10) will nearly equal to $5.97 - \ln(\beta_{rep.events})$, then $\beta_{rep.events} = \exp(-4) = 0.018$, it' s not consistent with experts opinion of $\beta_{rep.events}$, besides, standard deviation $\sigma_0 = 2$ is not consistent with standard deviation (which is 5) from experts, either.

Similarly, if the given lognormal prior for ϕ is sensitive, then I derived x should be around 3.87, this value is not consistent with expert' s opinion of x. Therefore $\phi \sim Lognormal(\mu_0, \sigma_0^2)$ isn' t an appropriate prior.

Simulation: Besides, I carried out a simulation by rjags, converged simulation results are as follow:

- The probability of value of $\log(\phi)$ less than -4 or greater than 4: 0.52
- The probability of absolute value of $\beta_{rep.events}$ less than $\log(4)/5$: 0.39

Simulation results are consistent with computations, indicates that $\phi \sim Lognormal(\mu_0, \sigma_0^2)$ isn' t an appropriate prior.

01-(e): Tuning model by adding extra variance

Expand the previous model in (1b) by including an extra variance term θ. I set the following generalized linear model with a logarithm link connecting the response mean with the linear combination of the covariates and θ:

$$\log(\exp(-\theta)\mu_i) = \beta_0 + \beta_1 \widetilde{Rep.event_i} + \beta_2 EADS1_i + \beta_3 EADS2_i \qquad (2)$$

$$\mu_i = \exp\{\beta_0 + \beta_1 \widetilde{Rep.event_i} + \beta_2 EADS1_i + \beta_3 EADS2_i + \theta\}$$

where θ represent the high variability in the number of deaths for different years.

After running JAGS, similar with convergence diagnosis in 01-(b), Brooks-Gelman-Rubin statistic of β and θ are all around 1 and there' s no trend in their trace plots, which indicates MCMC converged. Comparisons are as follow:

- $E_{model.e}[\exp\{\beta_0\}]$=2.58 is the expected number of deaths for a year between 1986 and 1993. $E_{model.b}[\exp\{\beta_0\}]$=3.45 is the expected number of deaths for a year between 1986 and 1993.
- A year with a reported events one unit more than the mean has an expected number of deaths increased of $E_{model.e}[\exp\{\beta_1\}]$=1.26, $E_{model.b}[\exp\{\beta_1\}]$ =1.22, while the other variables are hold constant.
- A year between 1994-2003 has an expected number of deaths increased of $E_{model.e}[\exp\{\beta_2\}]$ =1.04, $E_{model.b}[\exp\{\beta_2\}]$= 0.91, while the other variables are hold constant.
- A year after 2004 has an expected number of deaths increased of $E_{model.e}[\exp\{\beta_3\}]$ = 0.53, $E_{model.b}[\exp\{\beta_3\}]$ = 0.41, while the other variables are hold constant.

01-(f):Compare original and tuned models

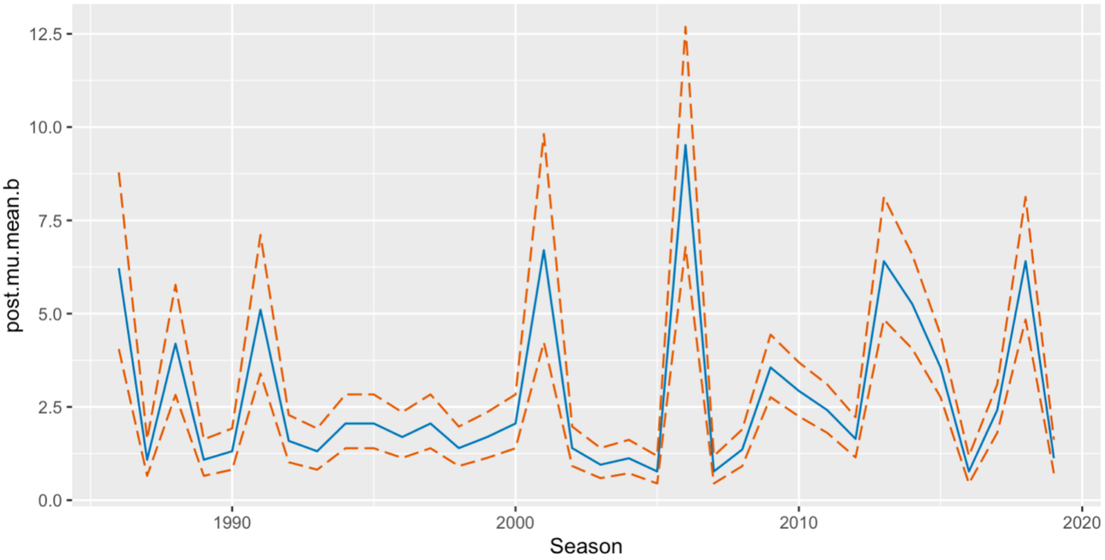


Figure 4: posterior means with 90% point-wise credible intervals for model in (b)

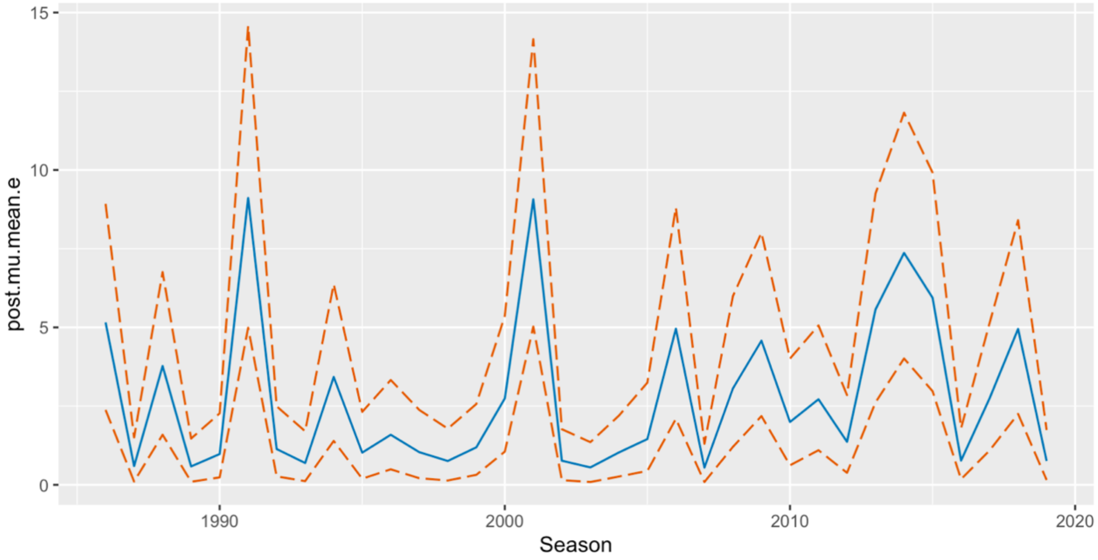


Figure 5: posterior means with 90% point-wise credible intervals for model in (e)

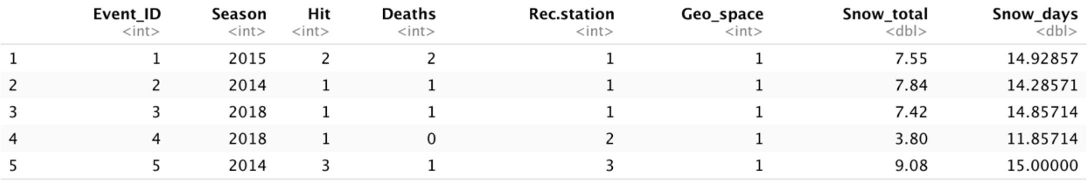
Deviance Information Criterion of model in (b) is 137.9, model in (e) is 110, difference between them is 13.34, Therefore, based on DIC the model in (e) is better.

Plots in (a) and (f) indicates that model in (e) captures more fluctuation and predicted death numbers are more accurate. Death number in Figure 4 seems more similar to Figure 2 (Actual Deaths), especially during 2004-2019. For example, in 2005, Figure 3 shows that deaths rush up, but Figure 4 doesn’ t, which is more consistent with the actual death number. However, 90% credit interval band of Figure 4 is wider than Figure 3, this is because the introduction of θ in model (e).

Sum up what are discussed afore, model in (e) is better from my point of view.

Problem 02

02-(a):Exploratory on avalanche data part 2



	Event_ID	Season	Hit	Deaths	Rec.station	Geo_space	Snow_total	Snow_days
	<int>	<int>	<int>	<int>	<int>	<int>	<dbl>	<dbl>
1	1	2015	2	2	1	1	7.55	14.92857
2	2	2014	1	1	1	1	7.84	14.28571
3	3	2018	1	1	1	1	7.42	14.85714
4	4	2018	1	0	2	1	3.80	11.85714
5	5	2014	3	1	3	1	9.08	15.00000

Table 2: Avalanches part 2 data frame head

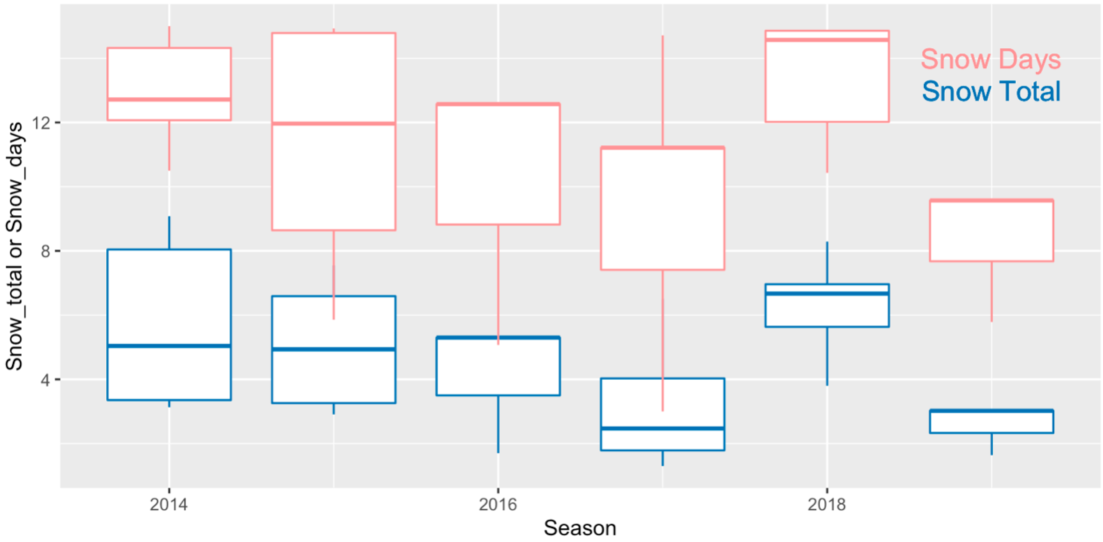


Figure 6: Boxplot for Snow_days and Snow_total

```
            Season  Snow_total  Snow_days
Season         1.0        -0.1       -0.1
Snow_total    -0.1         1.0        0.8
Snow_days     -0.1         0.8        1.0
```

Table 3: Correlation matrix of Season, Snow_total and Snow_days

Correlation between Snow_days and Snow_total is 0.8, which indicates there' s collinearity if they are both in a general linearly combined model.

02-(b) : fitting a bayes logistic model

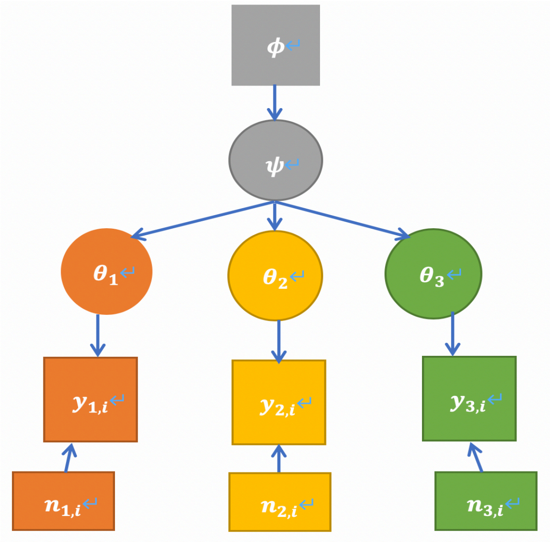


Figure 7: DAG of Hierarchical Model in 2-(b)

Hierarchical formulas are as follow:

$$y_{j,i}|\theta_i \sim Binomial(\theta_i, n_i)$$

$$log\left(\frac{\theta_i}{1-\theta_i}\right) = \psi_i + \beta_1 \times \widetilde{Season_\iota} + \beta_2 \times \widetilde{Snow\_total_\iota} + \beta_3 \times \widetilde{Snow\_day_\iota}$$

$$\psi \sim Normal(0, \sigma^2)$$

Where $\widetilde{Season_i}$, $\widetilde{Snow\_total_i}$ and $\widetilde{Snow\_day_i}$ are centered covariates, ψ_i is the i^{th} category of Geo_space corresponding random effect, n_i is the number of hits, θ_i is the probability of deaths. After running rJAGS, Brooks-Gelman-Rubin statistic of β and ψ_i are all around 1 and there’ s no trend in trace plots, which indicates MCMCs are converged. But β_2 and β_3 ACF plots are lagged. Results interpretations are as follow:

An event with year one unit more than the mean has an expected log odds of death probability changed of $E[\beta_1] = -0.12$, while the other variables are hold constant.

An event with Snow_total one unit more than the mean has an expected log odds of death probability changed of $E[\beta_2] = -0.15$, while the other variables are hold constant.

An event with year one unit more than the mean has an expected log odds of death probability changed of $E[\beta_3] = 0.04$, while the other variables are hold constant.

$E[\psi_1] = 0.028$, $E[\psi_2] = -0.38$, $E[\psi_3] = -1$ are expected log odds of death probability with random effect (corresponding to three different Geo_space) when Season, Snow_total, Snow_day are equal to their mean.

02-(c):rJAGS simulation on model parameters estimation

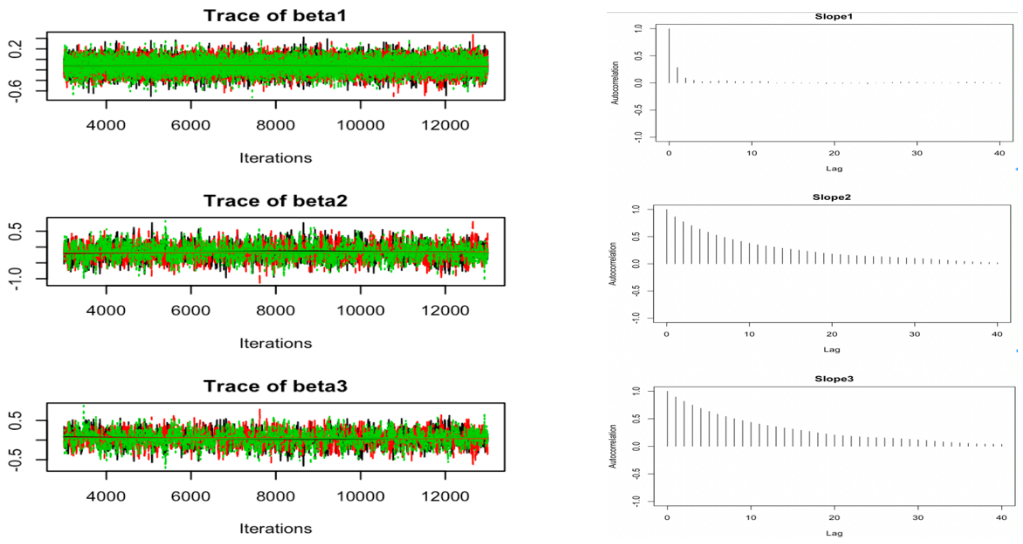


Figure 8: Trace plots and Gelman Plots in Hierarchical model in 02-(b)

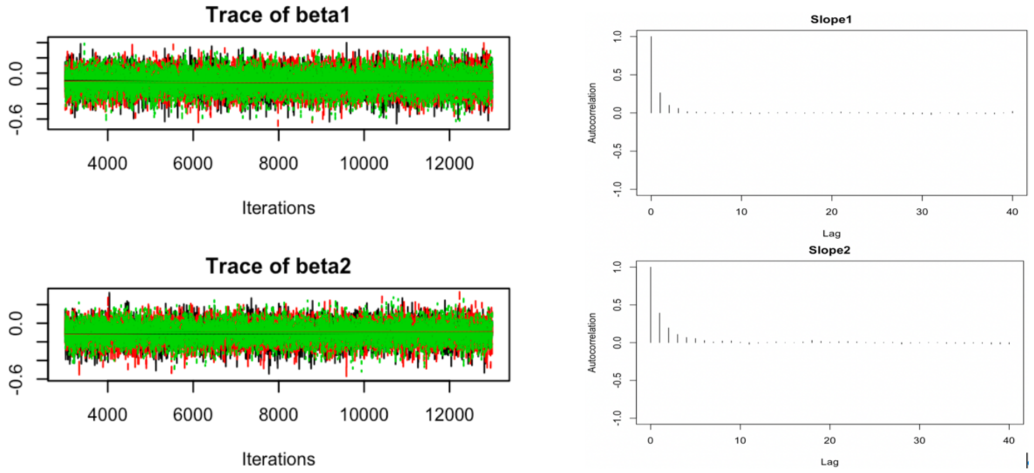


Figure 9: Trace plots and Gelman Plots in Hierarchical model in 02-(b)

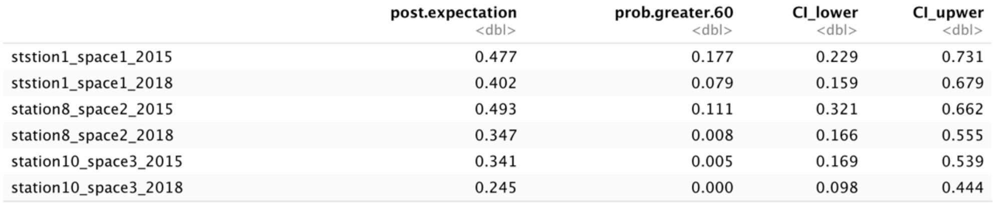
Trace plots of β_2 and β_3 are not as completely random as trace plot of β_1 in Hierarchical model in 02-(b), and their corresponding ACF plots are obviously lagged. But trace plots of slopes in model in 02-(c) are random and ACF is not lagged, which means MCMCs in model in 02-(c) converge better. This is because Snow_Total and Snow_Days are highly correlated. Model in 02-(c) fixed this collinearity problem by delete Snow_Days. Results interpretations from model in 02-(c) are as follow:

An event with year one unit more than the mean has an expected log odds of death probability changed of $E[\beta_1] = -0.11$, while the other variables are hold constant.

An event with Snow_total one unit more than the mean has an expected log odds of death probability changed of $E[\beta_2] = -0.11$, while the other variables are hold constant. $E[\beta_1]$ and $E[\beta_2]$ aren’ t changed a lot, this is also because Snow_Days and Snow_Total are highly correlated, Snow_Day doesn’ t provide new “information” to model. $E[\psi_1] = 0.023$, $E[\psi_2] = -0.35$, $E[\psi_3] = -1.05$, they are similar to results in 02-(c).

02-(d): Several key posterior probabilities

Based on model obtained in (c), I estimated the posterior expected value and 95% credible interval of the proportion of deaths near recording stations 1, 8 and 10 for the Seasons 2015 and 2018. Specific values are given in Table 4.



	post.expectation	prob.greater.60	CI_lower	CI_upwer
	<dbl>	<dbl>	<dbl>	<dbl>
ststion1_space1_2015	0.477	0.177	0.229	0.731
ststion1_space1_2018	0.402	0.079	0.159	0.679
station8_space2_2015	0.493	0.111	0.321	0.662
station8_space2_2018	0.347	0.008	0.166	0.555
station10_space3_2015	0.341	0.005	0.169	0.539
station10_space3_2018	0.245	0.000	0.098	0.444

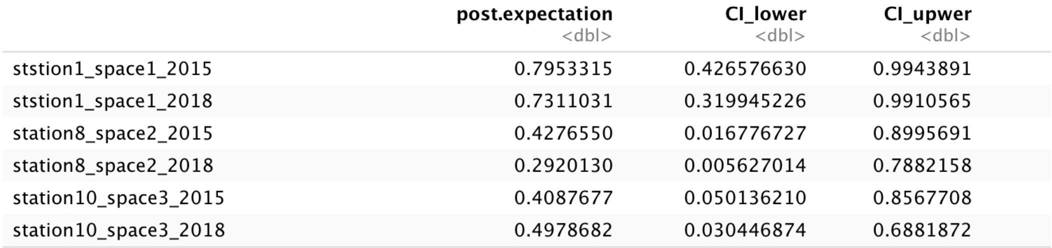
Table 4: posterior expected value and 95% credit interval 2(d)

According to Table 4, proportion of death decrease as year increase (while other covariates remain the same). Besides, proportion of death in space 3 is lower than space 1 and 2; space1 and 2 have similar proportion of death (while other covariates remain the same).

Compare the probability of a proportion of deaths greater than 60%: station 10< station 1 < station 8, Year 2018<Year 2015.

02-(e): Several key posterior expectations

Set up a new hierarchical model where the random effects are placed on the recording stations (Rec.station), after diagnosis, all parameters and random effects converge well. DIC of model in 2(c) is 87.9, model in 2(d) is 75.36, based on DIC, model in 2(d) performs better.



	post.expectation	CI_lower	CI_upwer
	<dbl>	<dbl>	<dbl>
ststion1_space1_2015	0.7953315	0.426576630	0.9943891
ststion1_space1_2018	0.7311031	0.319945226	0.9910565
station8_space2_2015	0.4276550	0.016776727	0.8995691
station8_space2_2018	0.2920130	0.005627014	0.7882158
station10_space3_2015	0.4087677	0.050136210	0.8567708
station10_space3_2018	0.4978682	0.030446874	0.6881872

Table 5: posterior expected value and 95% credit interval 2(e)

Proportions of death (posterior expected value) increase because model in 2(d) provides a better fit. And lengths of credit intervals are longer compared with credit intervals in Table 4, this is probably because place random effects on a more specific category variable increase variability of model,

02-(f): ideas of improving model

In order to capture variability in the different recording stations and mountain areas in one single hierarchical model, we can employ another random effect based on model in 2(e).

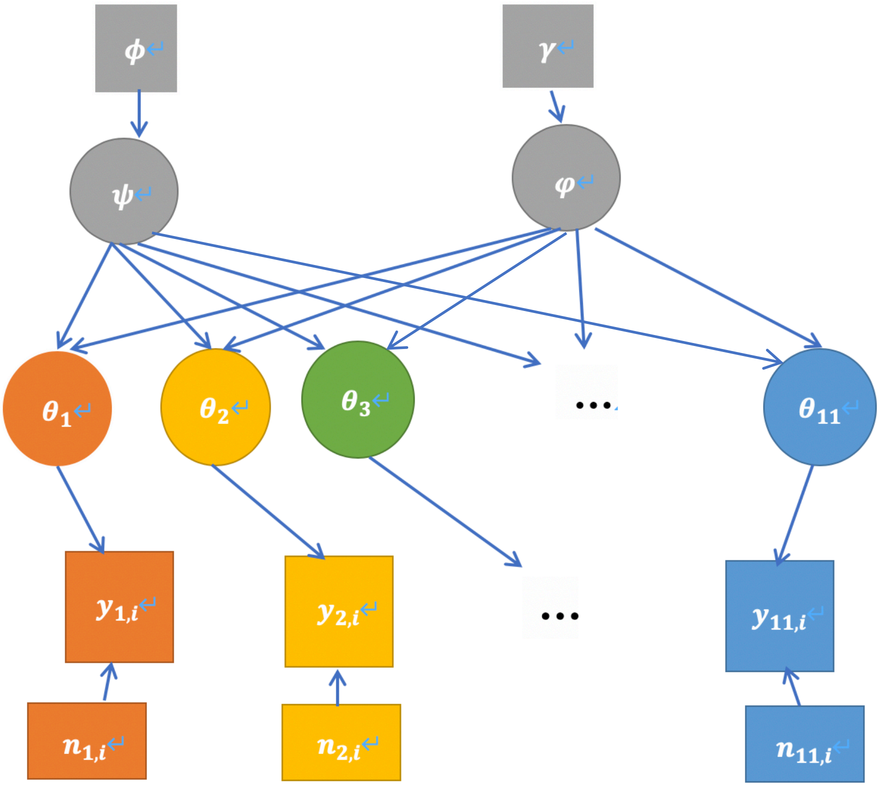


Figure 10: DAG of Hierarchical Model in 2-(f)

Hierarchical formulas are as follow:

$$y_{j,i}|\theta_i \sim Binomial(\theta_i, n_i)$$

$$log\left(\frac{\theta_i}{1-\theta_i}\right) = \psi_{l,i} + \varphi_{k,i} + \beta_1 \times \widetilde{Season_\imath} + \beta_2 \times \widetilde{Snow\_total_\imath} + \beta_3 \times \widetilde{Snow\_day_\imath}$$

$$\psi \sim Normal(\mu_1, \sigma_1^2)$$
$$\varphi \sim Normal(\mu_2, \sigma_2^2)$$

Where ψ_l is the l^{th} category of Geo_space corresponding random effect, φ_k is the k^{th} category of Rec.station corresponding random effect, k = 1,2⋯,10,11. Each θ_i has two random effects, one from ψ, the other from φ. Other symbols meanings are similar to DAG in 2(b)