Project 3: Bayesian Modeling of Hurricane Trajectories

Wenhao Gou, Qetsiyah Wang, Jared Klug, Michael Yan

Introduction

«««< HEAD Hierarchical Bayesian Model is the combination of two methods: Linear Regression Model in Hierarchical Form and Bayesian Inference methods. =======

Hierarchical Bayesian Model is the combination of two methods: Linear Regression Model in Hierarchical Form and Bayesian Inference methods. »»»> cba9289fa7670dcb898557df6cfbcb0e565526d4

Hierarchical form model is the linear regression model containing both Within-group analysis and Between-group analysis. Observations for each subject, from hierarchical form model, are usually collected with natural heterogeneity across the whole population over research time periods. This natural heterogeneity can be regarded as subject-specific mean response trajectories (i,e,. random effects) for each individual group. Nested data for each individual group violates the independent assumption for the linear regression model. Holding these individual-specific effects, the overall mean response over time across the whole research population is still considered as linearly over time, which is called Population-level effects (i.e,. fixed effects).

$$\underbrace{\pi(\theta|X)}_{posterior\ distribution} \propto \underbrace{\pi(X|\theta)}_{likelihood} \times \underbrace{\pi(\theta)}_{prior\ distribution}$$

Bayesian Inference is a statistical inference method about parameter. Before data collection, a proper prior distribution of parameter θ is set based on our belief about θ. Then after data collection $\mathbf{X} = (X_1, X_2, ..., X_n)$, the belief of parameter θ would be updated by exploring the posterior distribution of θ based on observed data and its pre-assumed likelihood function $L(X;\theta)$. The linear regression model in hierarchical form incorporating with Bayesian inference is implemented with Markov Chain Monte Carlo Integration algorithm for updating the parameter estimation in the final MCMC stationary phase.

Objective

In this project, we are going to explore whether the population-level changing trend of hurricane wind speed over years through a total of 356 groups of hurricanes. Each hurricane contained its own individual-level-specific effects. The hierarchical Bayesian model for the ith hurricane is shown as

$$Y_i(t{+}6) = \beta_{0,i}{+}X_{i_{Month}}\beta_{1,i}{+}X_{i_{Year}}\beta_{2,i}{+}X_{i_{Type}}\beta_{3,i}{+}X_{i_{Y_{i,t}(t)}}\beta_{4,i,t}{+}\Delta_{i,t_{lat}}\beta_{5,i,t}{+}\Delta_{i,t_{lon}}\beta_{6,i,t}{+}\Delta_{i,t_{Speed}}\beta_{7,i,t}{+}\epsilon_i(t),$$

where $i = 1,\ 2, ..., m$ standing for each hurricane group and $t = 1,\ 2, ..., t_i$ standing for each recorded time point within ith hurricane group.

The provided hierarchical Bayesian model for hurricane trajectories for ith hurricane contains 4 population-level effects (i.e,. fixed): $X_{i_{Month}}$ - the month of year when hurricane started, $X_{i_{Year}}$ - the calender year of the hurricane, $X_{i_{Type}}$ - the type of hurricane, and 4 individual-level- effects (i.e,. random): $X_{i_{Y_{i,ti}(t)}}$ - the ith wind speed at $t-1$ time point for t time point, $\Delta_{i,t_{lat}}$, $\Delta_{i,t_{lon}}$, $\Delta_{i,t_{Speed}}$ - the change of latitudes, longitudes and wind speeds between two recorded time points. 4 prior information are provided as following:

$$\boldsymbol{\beta_i} \sim N(\boldsymbol{\beta}, \Sigma^{-1}),\ \pi(\sigma^2) \propto \frac{1}{\sigma^2},$$

$$\pi(\boldsymbol{\beta}) \propto 1,\ \pi(\boldsymbol{\Sigma}^{-1}) \propto |\boldsymbol{\Sigma}|^{-(d+1)/2} exp(-\frac{1}{2}\boldsymbol{\Sigma}^{-1}),$$

Then our Hierarchical Bayesian Model for analyzing the trend of hurricane wind speed across years with considering all other covariates would be performed firstly by exploring estimated parameters from posterior distributions based on Bayes Theorem.

Hurricane Data

hurricane356.csv collected the track data of 356 hurricanes in the North Atlantic area since 1989. For all the storms, their location (longitude & latitude) and maximum wind speed were recorded every 6 hours. The data includes the following variables

1. **ID**: ID of the hurricanes
2. **Season**: In which **year** the hurricane occurred
3. **Month**: In which **month** the hurricane occurred
4. **Nature**: Nature of the hurricane

- ET: Extra Tropical
- DS: Disturbance
- NR: Not Rated
- SS: Sub Tropical
- TS: Tropical Storm

5. **time**: dates and time of the record

6. **Latitude** and **Longitude**: The location of a hurricane check point
7. **Wind.kt** Maximum wind speed (in Knot) at each check point

Posterior Distribution Inference and MCMC Methods

Posterior Distribution Inference

We have $\beta_i = (\beta_{0,i}, \beta_{1,i}, ..., \beta_{7,i,t})$ associated with the ith hurricane. According to the knowledge of Multivariate Linear Regression Model, we know that $Y_{i,t} \sim N(X_{i,t}\beta_i,\ \sigma^2 I_{t_i})$, denoting, $\boldsymbol{\mu}_{i,t} = \mathbf{X}_{i,t}\boldsymbol{\beta_i}$,

Then we have,

$$\pi(\mathbf{Y}_{i,t}|\boldsymbol{\mu}_{i,t}, \sigma^2, \boldsymbol{\beta}, \boldsymbol{\Sigma}^{-1}) \propto \frac{1}{\sqrt{\sigma^2}} exp\{-\frac{1}{2}(\mathbf{Y}_{i,t} - \boldsymbol{\mu}_{i,t})^T(\sigma^2 I_{t_i})^{-1}(\mathbf{Y}_{i,t} - \boldsymbol{\mu}_{i,t})\}$$

1. For $\pi(\boldsymbol{\beta_i}|.)$:

$$\pi(\boldsymbol{\beta_i}|.) \propto f(\boldsymbol{Y_i}|\boldsymbol{\beta_i}, \sigma^2) f(\boldsymbol{\beta_i}|\boldsymbol{\beta}, \boldsymbol{\Sigma}^{-1}) \tag{1}$$

$$\propto \left(\prod_{t=1}^{t_i} \frac{1}{\sqrt{\sigma^2}} exp\{-\frac{1}{2}(\mathbf{Y}_{i,t} - \boldsymbol{\mu}_{i,t})^T(\sigma^2 I_{t_i})^{-1}(\mathbf{Y}_{i,t} - \boldsymbol{\mu}_{i,t})\}\right)\left(|\boldsymbol{\Sigma}|^{-1/2} \exp\{-\frac{1}{2}(\boldsymbol{\beta_i} - \boldsymbol{\beta})^T \boldsymbol{\Sigma^{-1}}(\boldsymbol{\beta_i} - \boldsymbol{\beta})\}\right) \tag{2}$$

$$\propto \exp\{-\frac{1}{2}\left((\boldsymbol{Y_i} - \boldsymbol{X_i}\boldsymbol{\beta_i})^T(\sigma^{-2}\boldsymbol{I}_{t_i \times t_i})(\boldsymbol{Y_i} - \boldsymbol{X_i}\boldsymbol{\beta_i}) + (\boldsymbol{\beta_i} - \boldsymbol{\beta})^T \boldsymbol{\Sigma}_{8\times 8}^{-1}(\boldsymbol{\beta_i} - \boldsymbol{\beta})\right)\} \tag{3}$$

For the exponential term:

$$
\begin{aligned}
&(\boldsymbol{Y_i}-\boldsymbol{X_i}\boldsymbol{\beta_i})^T(\sigma^{-2}\boldsymbol{I}_{t_i\times t_i})(\boldsymbol{Y_i}-\boldsymbol{X_i}\boldsymbol{\beta_i})+(\boldsymbol{\beta_i}-\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}_{8\times 8}(\boldsymbol{\beta_i}-\boldsymbol{\beta}) && (4)\\
&=\boldsymbol{Y_i}^T\sigma^{-2}\boldsymbol{I}\boldsymbol{Y_i}^T+\boldsymbol{\beta_i}^T\boldsymbol{X_i}^T\sigma^{-2}\boldsymbol{I}\boldsymbol{X_i}\boldsymbol{\beta_i}-2\boldsymbol{Y_i}^T\sigma^{-2}\boldsymbol{I}\boldsymbol{X_i}\boldsymbol{\beta_i} && (5)\\
&\quad+\boldsymbol{\beta_i}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta_i}+\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}-2\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta_i} && (6)\\
&=\boldsymbol{Y_i}^T\sigma^{-2}\boldsymbol{I}\boldsymbol{Y_i}^T+\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}+\boldsymbol{\beta_i}^T(\boldsymbol{\Sigma}^{-1}+\boldsymbol{X_i}^T\sigma^{-2}\boldsymbol{I}\boldsymbol{X_i})\boldsymbol{\beta_i} && (7)\\
&\quad-2(\boldsymbol{Y_i}^T\sigma^{-2}\boldsymbol{I}\boldsymbol{X_i}+\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1})\boldsymbol{\beta_i} && (8)\\
&=\boldsymbol{R}+\boldsymbol{\beta_i}^T\boldsymbol{V}\boldsymbol{\beta_i}-2\boldsymbol{M}\boldsymbol{\beta_i} && (9)
\end{aligned}
$$

Where:

$$
\begin{aligned}
\boldsymbol{R}&=\boldsymbol{Y_i}^T\sigma^{-2}\boldsymbol{I}_{t_i\times t_i}\boldsymbol{Y_i}+\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta} && (10)\\
\boldsymbol{V}&=\boldsymbol{\Sigma}^{-1}+\sigma^{-2}\boldsymbol{X_i}^T\boldsymbol{X_i} && (11)\\
\boldsymbol{M}&=\sigma^{-2}\boldsymbol{X_i}^T\boldsymbol{Y_i}+\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta} && (12)
\end{aligned}
$$

Then, the exponential term can be reduced to:

$$(\boldsymbol{\beta_i}-\boldsymbol{V^{-1}M})^T\boldsymbol{V}(\boldsymbol{\beta_i}-\boldsymbol{V^{-1}M})-\boldsymbol{M}^T\boldsymbol{V^{-1}}\boldsymbol{M}^T+\boldsymbol{R}$$

We can ignore the latter 2 term as it is not related to $\boldsymbol{\beta_i}$. That indicate:

$$\pi(\boldsymbol{\beta_i}|.)\sim N(\boldsymbol{V^{-1}M},\boldsymbol{V^{-1}})$$

2.For $\pi(\sigma^2|.)$:

$$
\begin{aligned}
\pi(\sigma^2|.)&\propto f(\boldsymbol{Y}|\boldsymbol{\beta_i},\sigma^2)\cdot\pi(\sigma^2) && (13)\\
&\propto\left(\prod_{i=1}^{n}\prod_{t=1}^{t_i}\sigma^{-1}\exp\{-\frac{(\boldsymbol{Y}_{i,t}-\boldsymbol{\mu}_{i,t})^2}{2\sigma^2}\}\right)\frac{1}{\sigma^2} && (14)\\
&\propto(\sigma^2)^{-1-\frac{\sum t_i}{2}}\prod_{i=1}^{n}\prod_{t=1}^{t_i}\exp\{-\frac{(\boldsymbol{Y}_{i,t}-\boldsymbol{\mu}_{i,t})^2}{2\sigma^2}\} && (15)\\
&\propto\sigma^{-2-\sum t_i}\exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{t=1}^{ti}(\boldsymbol{Y}_{i,t}-\boldsymbol{\mu}_{i,t})^2\} && (16)
\end{aligned}
$$

So:

$$\sigma^2\sim\text{Inverse Gamma}(\frac{1}{2}\sum_{i=1}^{n}t_i,\frac{1}{2}\sum_{i=1}^{n}\sum_{t=1}^{ti}(\boldsymbol{Y}_{i,t}-\boldsymbol{\mu}_{i,t})^2)$$

3. For $\pi(\boldsymbol{\Sigma}^{-1}|.)$: We have the prior of $\pi(\boldsymbol{\Sigma^{-1}})\sim|\boldsymbol{\Sigma}|^{-(d+1)/2}exp(-\frac{1}{2}\boldsymbol{\Sigma}^{-1})$, which actually follows the Inverse Wishart distribution $(\boldsymbol{\Sigma^{-1}})\sim$ Inverse Wishart(v_0,S_0), where $v_0=0,S_0=1$

$$
\begin{aligned}
\pi(\boldsymbol{\Sigma}^{-1}|.)&\propto f(\boldsymbol{\beta_i}|\boldsymbol{\beta},\boldsymbol{\Sigma}^{-1})f(\boldsymbol{\Sigma}^{-1}) && (17)\\
&\propto\left(\prod_{i=1}^{n}\boldsymbol{\Sigma}^{-\frac{1}{2}}\exp\{-\frac{1}{2}(\boldsymbol{\beta_i}-\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta_i}-\boldsymbol{\beta})\}\right)|\boldsymbol{\Sigma}|^{-(d-1)/2}\exp\{-\frac{1}{2}\boldsymbol{\Sigma}^{-1}\} && (18)\\
&\propto|\boldsymbol{\Sigma}|^{-(n+d+1)/2}\exp\{-\frac{1}{2}tr(\boldsymbol{\Sigma^{-1}})-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{\beta_i}-\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta_i}-\boldsymbol{\beta})\} && (19)\\
&\propto|\boldsymbol{\Sigma}|^{-(n+d+1)/2}\exp\{-\frac{1}{2}tr\left(\boldsymbol{\Sigma^{-1}}(\boldsymbol{I}+\sum_{i=1}^{n}(\boldsymbol{\beta_i}-\boldsymbol{\beta})(\boldsymbol{\beta_i}-\boldsymbol{\beta})^T)\right)\} && (20)
\end{aligned}
$$

That indicate:

$$\boldsymbol{\Sigma}^{-1} \sim \text{Inverse Whishart}\left(n, \boldsymbol{I} + \sum_{i=1}^{n}(\boldsymbol{\beta_i} - \boldsymbol{\beta})^T(\boldsymbol{\beta_i} - \boldsymbol{\beta})\right)$$

4. For $\pi(\boldsymbol{\beta}|.)$:

$$\begin{aligned}
\pi(\boldsymbol{\beta}|.) &\propto f(\boldsymbol{\beta_i}|\boldsymbol{\beta}, \boldsymbol{\Sigma}^{-1})f(\boldsymbol{\beta}) && (21)\\
&\propto \left(\prod_{i=1}^{n}\exp\{-\frac{1}{2}(\boldsymbol{\beta_i} - \boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta_i} - \boldsymbol{\beta})\}\right) && (22)\\
&\propto \exp\{-\frac{1}{2}\left(\sum_{i=1}^{n}(\boldsymbol{\beta_i} - \boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta_i} - \boldsymbol{\beta})\right)\} && (23)\\
&\propto \exp\{-\frac{1}{2}\left(\sum_{i=1}^{n}\boldsymbol{\beta_i}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta_i} + \boldsymbol{\beta}^T n\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta} - \sum_{i=1}^{n}2\boldsymbol{\beta_i}^T\boldsymbol{\Sigma^{-1}}\boldsymbol{\beta}\right)\} && (24)
\end{aligned}$$

For the exponential term, if we set:

$$\begin{aligned}
\boldsymbol{V} &= n\boldsymbol{\Sigma}^{-1} && (25)\\
\boldsymbol{R} &= \sum_{i=1}^{n}\boldsymbol{\beta_i}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta_i} && (26)\\
\boldsymbol{M} &= \sum_{i=1}^{n}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta_i}\right) && (27)
\end{aligned}$$

Then, use the same technique when generating $\boldsymbol{\beta_i}$

$$R + \boldsymbol{\beta}\boldsymbol{V}\boldsymbol{\beta} - 2\boldsymbol{M}\boldsymbol{\beta} \propto (\boldsymbol{\beta} - \boldsymbol{V}^{-1}\boldsymbol{M})^T\boldsymbol{V}(\boldsymbol{\beta} - \boldsymbol{V}^{-1}\boldsymbol{M})$$

(NOTE: This is the same as using OLS to estimate $\boldsymbol{\beta}$ using all $\boldsymbol{\beta_i}$)

That indicate:

$$\boldsymbol{\beta} \sim N\left(Vec(\hat{\beta}_i), \boldsymbol{\Sigma}\ \beta_i^T\beta_i\right)$$

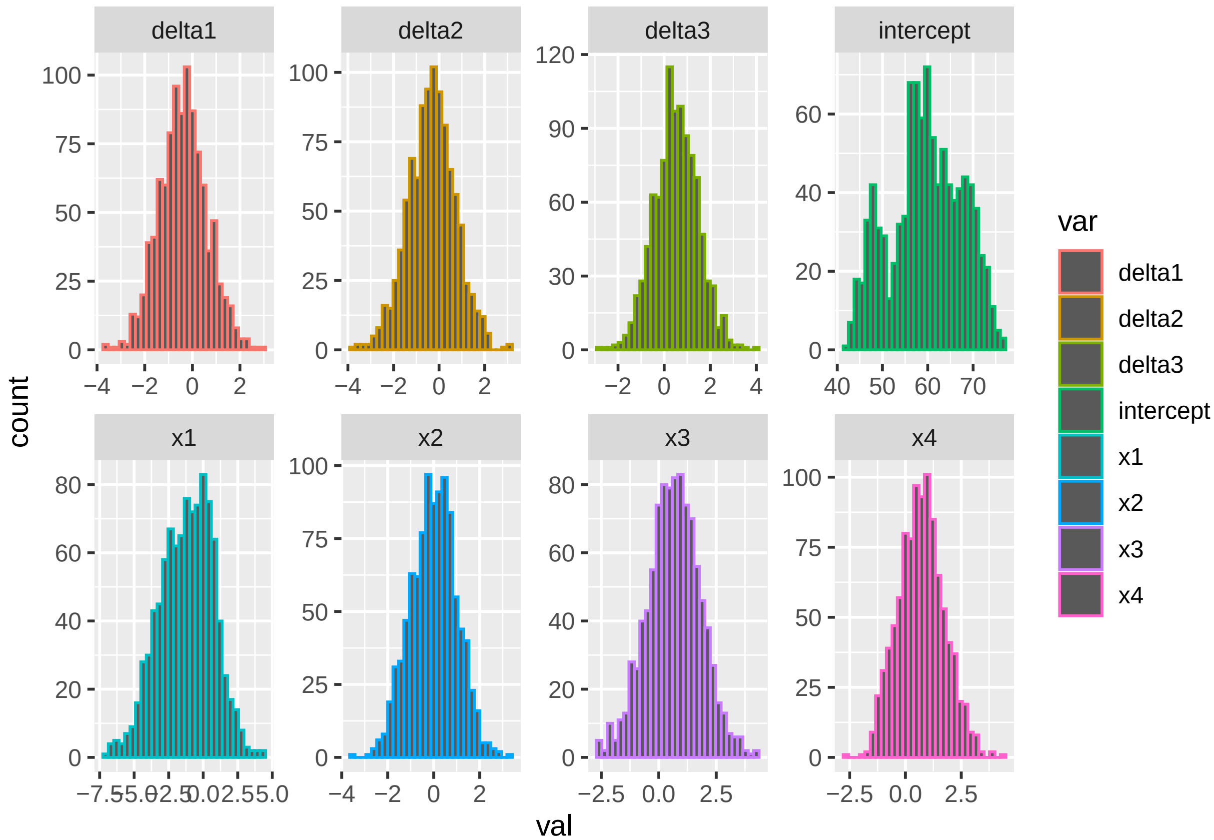
«««< HEAD

Results and dicussion

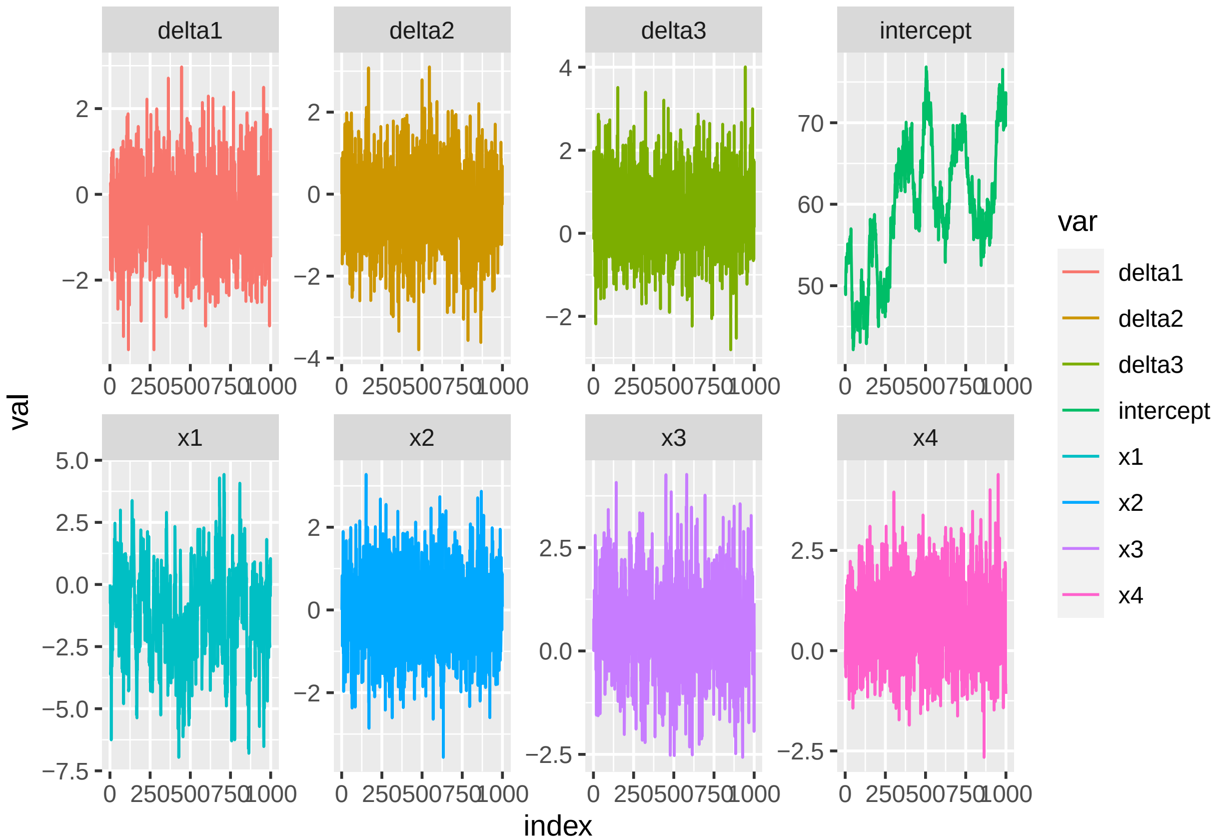
Test the chain with the first set of initial value

```
summary1 <- summaryplotsfun(test)
summary1[[1]]
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
summary1[[2]]
```

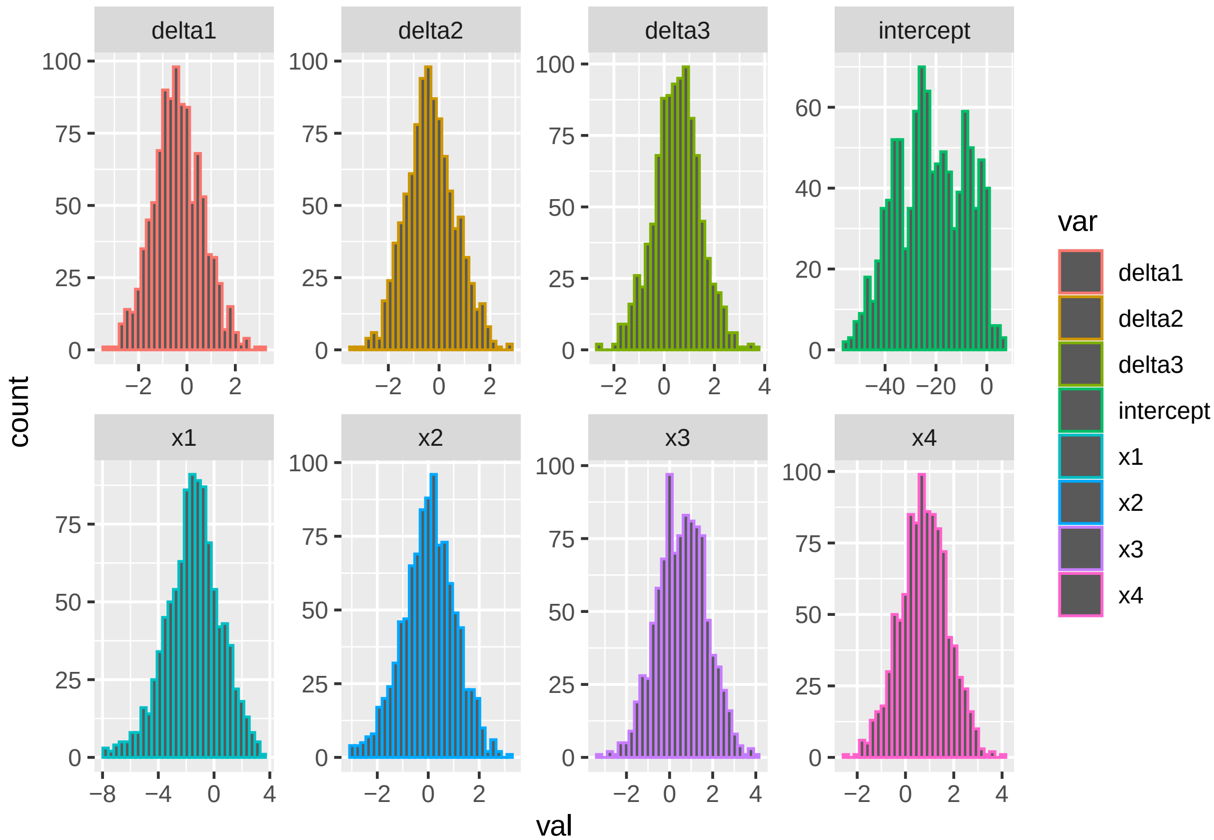


After testing the chain with the first set of initial values, from the density plots, we noticed that for all variables except the intercept, follow a normal distribution. From the line chart, we see that all variables converged except for the intercept. Therefore, we decided to do a warm start to see whether the problem is solved.

Test the chain with a warm start

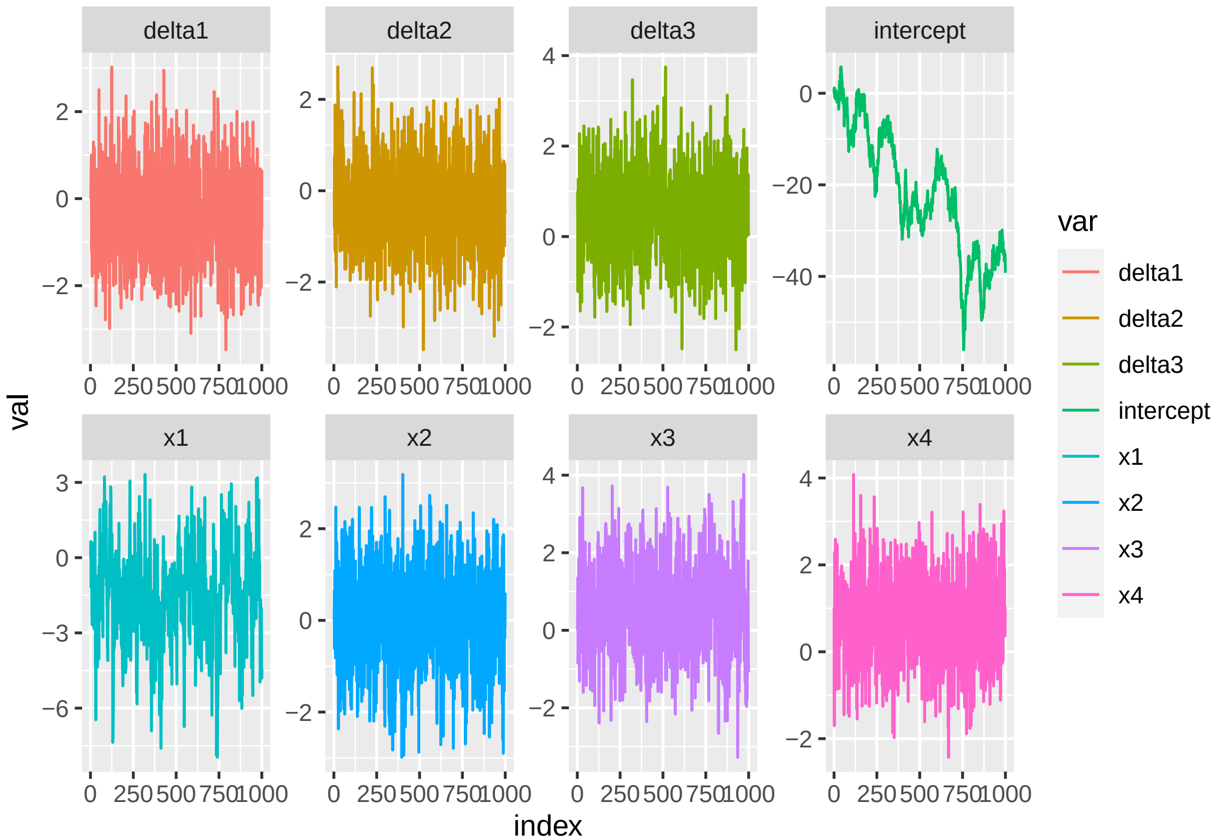
```r
summary2 <- summaryplotsfun(test2)
summary2[[1]]
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



delta1
delta2
delta3
intercept
x1
x2
x3
x4
count
val
var
delta1
delta2
delta3
intercept
x1
x2
x3
x4


```
summary2[[2]]
```



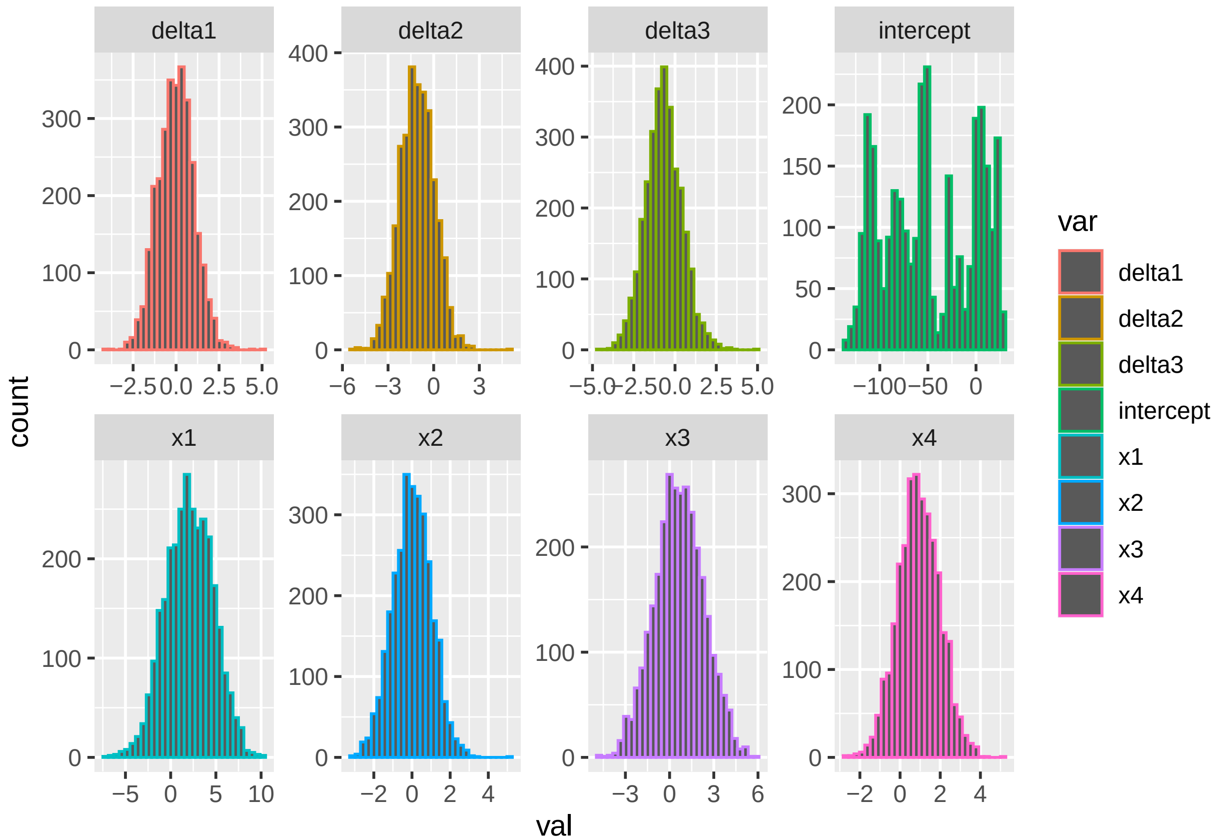
By doing warm start, for the intercept, it's density plot appeared to be a mixture of 2 normal distributions, suggesting that we need to subset the data. The line chart still shows no convergence. The nature of the wind speed is going up first and then going down, so we separate the data into up intervals and down intervals as one of our predictor, delta3.

Test the chain with subsetted data

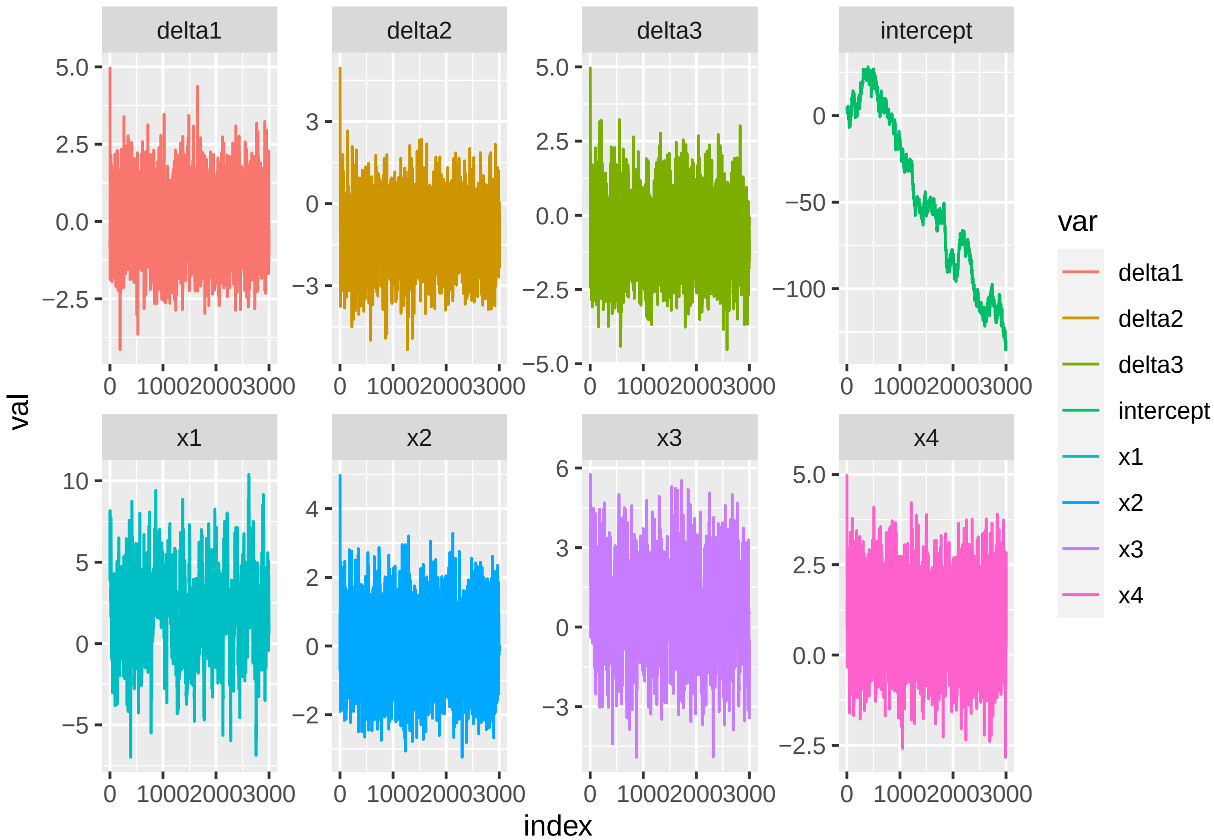
Plots for upchain

```
summaryup <- summaryplotsfun(upchain)
summaryup[[1]]
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



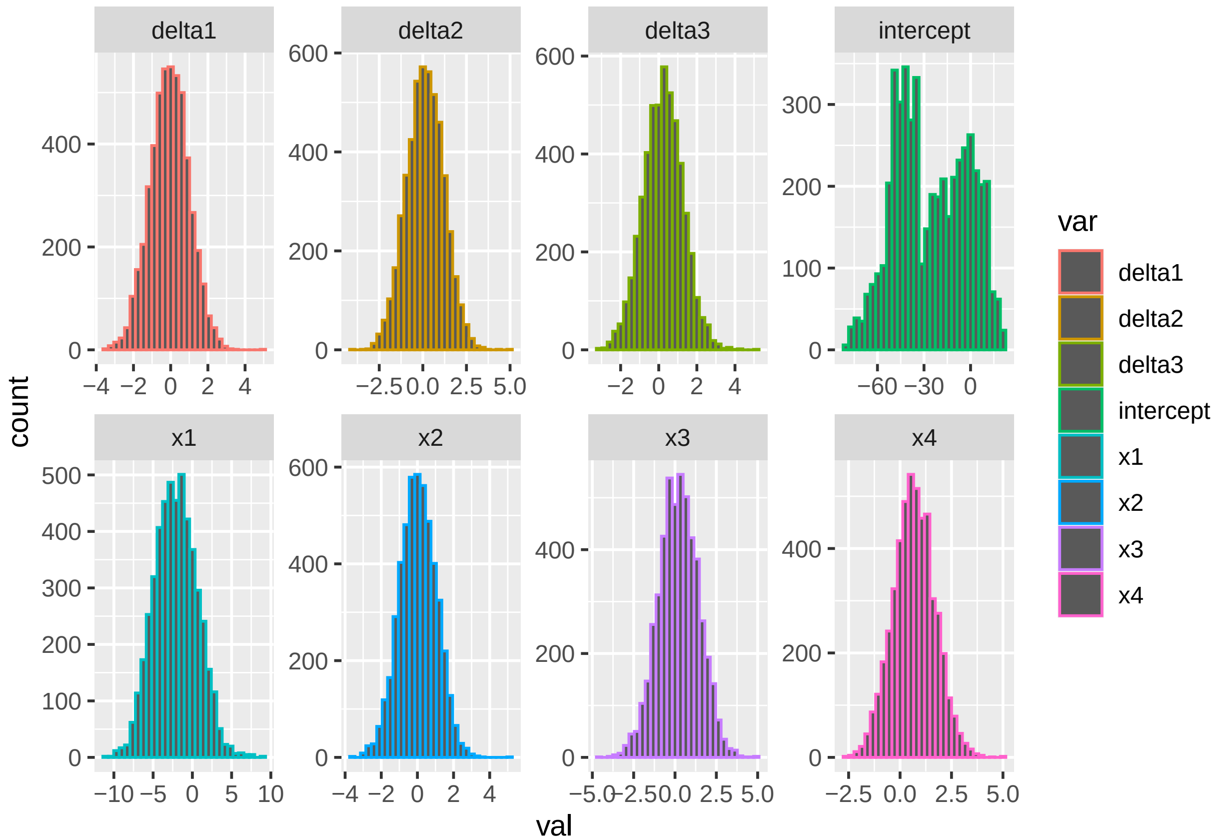

```
summaryup[[2]]
```



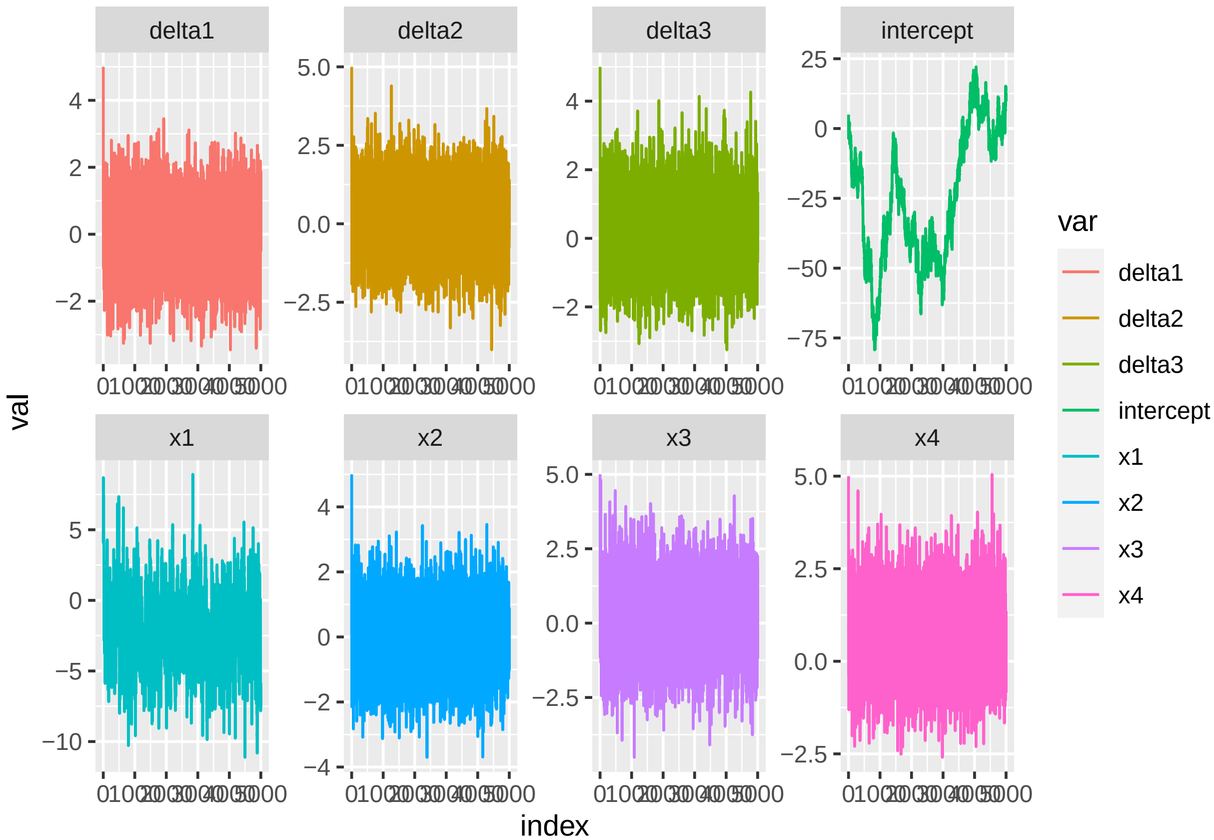
Plots for downchain

```
summarydown <- summaryplotsfun(downchain)
summarydown[[1]]
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
summarydown[[2]]
```



The “upchain” (wind speed increasing monotonically) converged in about 2000-3000 steps but the “downchain” did not seems to converge. This may due to a bad prior.

======= # Results »»»> cba9289fa7670dcb898557df6cfbcb0e565526d4