Analysis of Beijing PM2.5 dataset using MCMCglmm

Guanghui Min, Muzhe Guo, Naichen Ni

University of Michigan

guanghui, muzheguo, ncni @umich.edu

September 20, 2019

Overview

1. Overview of the model
   - Fixed vs Random Effects
   - Brief introduction to Linear Mixed Model
2. Preprocess the dataset
   - Overview of the dataset
   - Box-cox transformation for response
   - Choose preidictors
3. Apply the model on the dataset
   - Apply the simple linear model
   - Apply linear mixed model using MCMCglmm
   - Compare the mixed model to simple regression model

Fixed vs Random Effects

In the simple linear model, we assume that parameters are unknown constants.

- Regression: b is some unknown (constant) coefficient vector
- ANOVA: μ_j are some unknown (constant) means
- These are referred to as fixed effects.

Unlike fixed effects, random effects are NOT unknown constants.

- Random effects are random variables in the population
- Typically assume that random effects are zero-mean Gaussian
- Typically want to estimate the variance parameter(s)

Brief intro to LMM

The model can be represented as:

$$Y = X\beta + Zu + \epsilon$$

where $Y \in \mathbb{R}^m, \beta \in \mathbb{R}^n, X \in \mathbb{R}^{m\times n}, Z \in \mathbb{R}^{m\times k}$,
and $u \sim \mathcal{N}(0_k, G_k), \epsilon \sim \mathcal{N}(0_m, R_m)$

- $X\beta$ is referred as fixed effects and $Z\alpha$ is referred as the random effect part;
- Typically we assume α and ϵ are independent;
- X and Z are known design matrices relating the observations to y, β and u respectively.

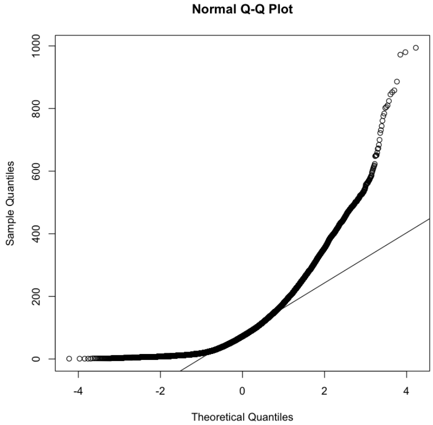
Beijing PM2.5 dataset

After deleting the missing values in the dataset, there are 41755 observations of 13 variables.

- year: year of data in this row
- month: month of data in this row
- day: day of data in this row
- hour: hour of data in this row
- pm2.5: PM2.5 concentration (ug/m^3)
- DEWP: Dew Point $(\hat{\alpha}_{,,} f)$
- TEMP: Temperature $(\hat{\alpha}_{,,} f)$
- PRES: Pressure (hPa)
- cbwd: Combined wind direction
- Iws: Cumulated wind speed (m/s)
- Is: Cumulated hours of snow
- Ir: Cumulated hours of rain

Box-cox transformation

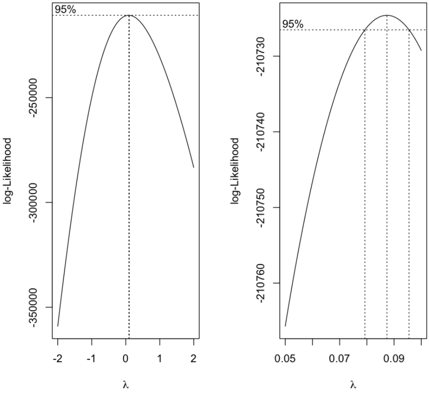
We choose the variable **pm2.5** as the response.



The response is severely skewed.

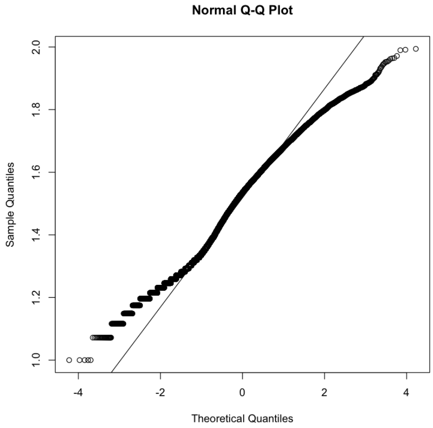
Box-cox transformation

We use box-cox plot for diagnostics:



For better interpretation, we choose $\lambda = 0.1$. Then $g_\lambda(y) = y^{0.1}$

Box-cox transformation

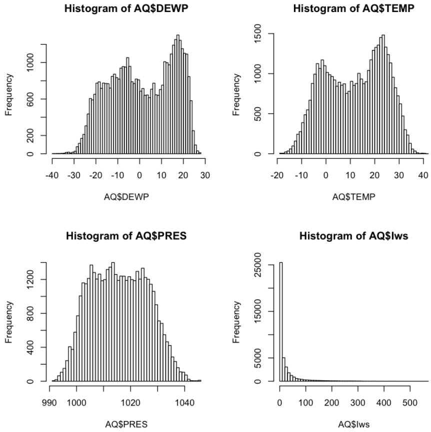


We can see that the transformed response follows a short-tailed distribution. For short-tailed distributions, the consequences of nonnormality are not serious and can reasonably be ignored.

Choose preidictors

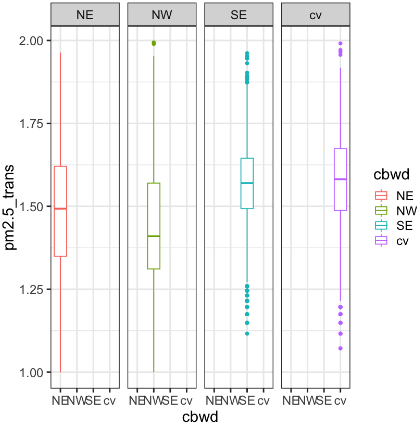
To avoid time series analysis, we tend to not use the variable **year**, month, day and **hour**. We create a new variable **season** to simply indicate the season of the row. What's more, as more than 95% of the entries of **IS**, **Ir** are zeros, we decide to deprecate them as well.
Finally, we will choose **DEWP**, **TEMP**, **PRES** and **Iws** as fixed effects and **cbwd** and **season** as random effects.

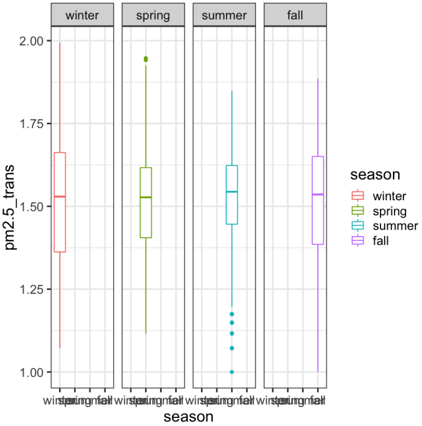
Choose preidictors



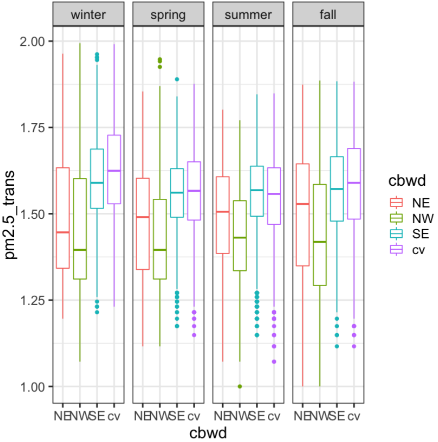
Choose preidictors

Here is the boxplots of transformed response grouped by **cbwd** and **season**.





Choose preidictors



Simple linear model

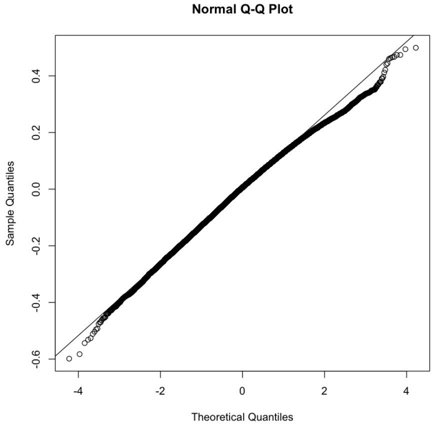
We first look the result of the simple model:

```
fit.linear=lm(pm2.5_trans~DEWP+TEMP+PRES+Iws,data=AQ)
summary(fit.linear)
```

```
Call:
lm(formula = pm2.5_trans ~ DEWP + TEMP + PRES + Iws, data = AQ)

Residuals:
     Min       1Q   Median       3Q      Max
-0.59857 -0.08534  0.00579  0.08956  0.49906

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.482e+00  1.149e-01   39.00   <2e-16 ***
DEWP         8.094e-03  8.306e-05   97.45   <2e-16 ***
TEMP        -1.016e-02  1.066e-04  -95.28   <2e-16 ***
PRES        -2.788e-03  1.124e-04  -24.82   <2e-16 ***
Iws         -6.635e-04  1.327e-05  -50.01   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1267 on 41750 degrees of freedom
Multiple R-squared:  0.3258,    Adjusted R-squared:  0.3258
F-statistic:  5045 on 4 and 41750 DF,  p-value: < 2.2e-16
```

Simple linear model

QQplot for residuals of the fit:



It turns out to be a good fit.

Using Non-informative prior

$$\begin{bmatrix} \epsilon \\ u \end{bmatrix} \sim \mathcal{N}(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} R & 0 \\ 0 & G \end{bmatrix})$$

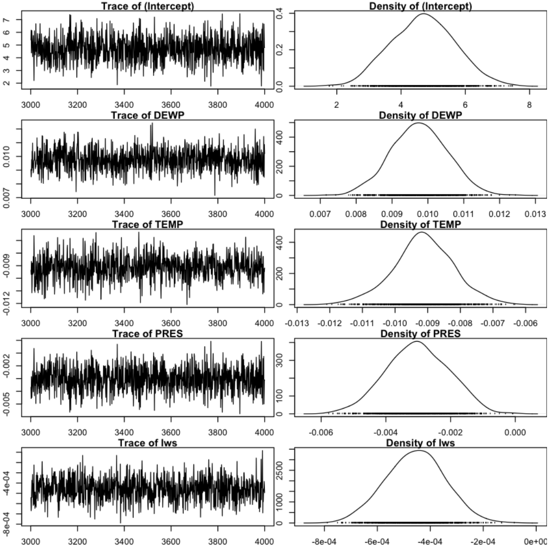
$$f(\epsilon, u|\beta, G, R) \propto |R|^{-1/2}|G|^{-1/2}$$

$$\times \exp(-\frac{1}{2}(y - X\beta - Zu)^T R^{-1}(y - X\beta - Zu) - \frac{1}{2}u^T G^{-1}u)$$

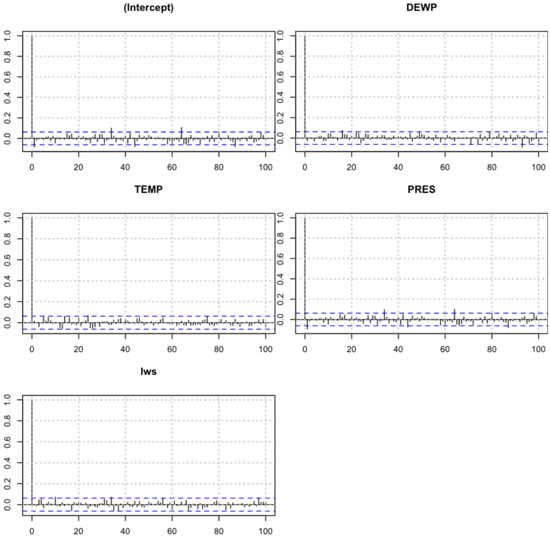
$$\hat{\beta}|y, u, G, R = (X^T V^{-1}X)^{-1}X^T V^{-1}y$$

where $V = ZGZ^T + R$ Then we assume the prior distribution of G, R follows inverse-Wishart distribution and apply gibbs sampler in this problem.

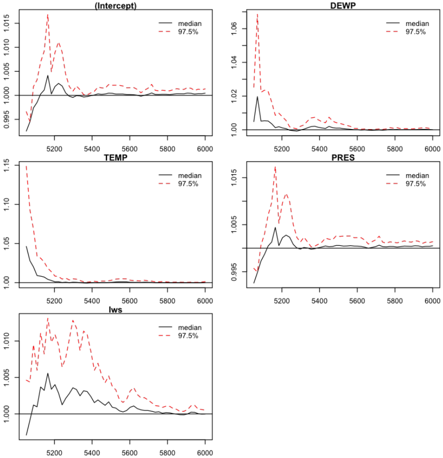
Using Non-informative prior



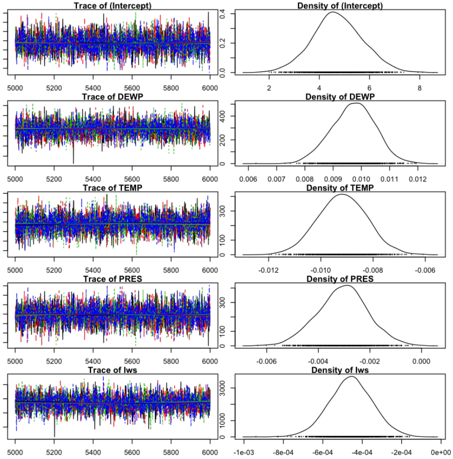
Using Non-informative prior



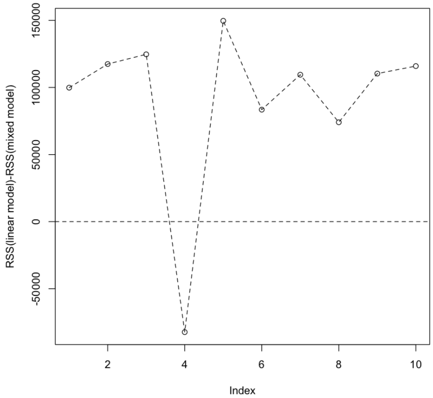
Using Non-informative prior



Using Non-informative prior



Compare the two methods



The advantage of the mixed model is that it does not assume the independency between the observations.

Thank you!