Data Analysis of the relationship between the Atmospheric Particulate Matter and It's Satellite Estimation

Country-by-country calibration of satellite data

1. Introduction

This report aims to investigate the relationship between the ground monitor measurements of particulate matter of less than 2.5 microns in diameter(PM 2.5) and a satellite data product that provides an estimate of PM 2.5. It is costful and unrealistic to build ground station every where to measure PM 2.5. Thus, it is helpful and important to build mmodels to check how well the satellite estimation can be.

2. Data

The first six rows of the data is as follows:

iso3	logPM25	logSAT	country_name_1	country_code_1	Super_region_name_1
AFG	4.454347	3.255458	Afghanistan	119	South Asia
AFG	4.219508	3.506863	Afghanistan	119	South Asia
ALB	2.772589	2.968772	Albania	72	Central Europe, Eastern Europe, Central Asia
AND	2.397895	1.888432	Andorra	61	High income
ARE	3.970292	3.821102	United Arab Emirates	118	North Africa / Middle East
ARE	4.143135	3.821267	United Arab Emirates	118	North Africa / Middle East

For each observation, the country iso, logPM25, logSAT, country_name, country_code, and Super_region_name are given. logPM25 represents the logarithm value of ground measured PM 2.5. logSAT represents the logarithm value of satellite estimation of PM 2.5. Iso, country_name and country_code represents the location of the ground station, while Super_region_name is the super region that country locates. To make it easier for calculation, a numeric level of Super_region_name is labelled and add as super_region_code.

super_region_name	super_region_code
South Asia	5
Central Europe, Eastern Europe, Central Asia	1
High income	2
North Africa / Middle East	4
Latin America and Caribbean	3
Southeast Asia, East Asia and Oceania	6
Sub-Saharan Africa	7

There is also an adjacency matrix shows the neighbouring country of each country.

3. Spatial Statistical Concepts

The aim is to estimate the PM 2.5 concentration in each country based on its log of satellite data. This goal can be gained through regression.

3.1. Parameter Estiamtion

The classic approach to estimate the parameter set $\theta = (\beta_0, \beta_1)$ is minimizing the Loss(Risk) funtion. However, it requires to have adequent amount of data. In the PM 2.5 case, the number of ground station is not likely to change or be adquent in every country. Thus enough data to do classic inference is unrealistic. The **Bayesian inference** is used instead. This is another approach to estimate the parameter that examines what values of θ are consistent with the PM 2.5 data. It does not require infinite amount of data, but focus on the data on hand.

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)}$$

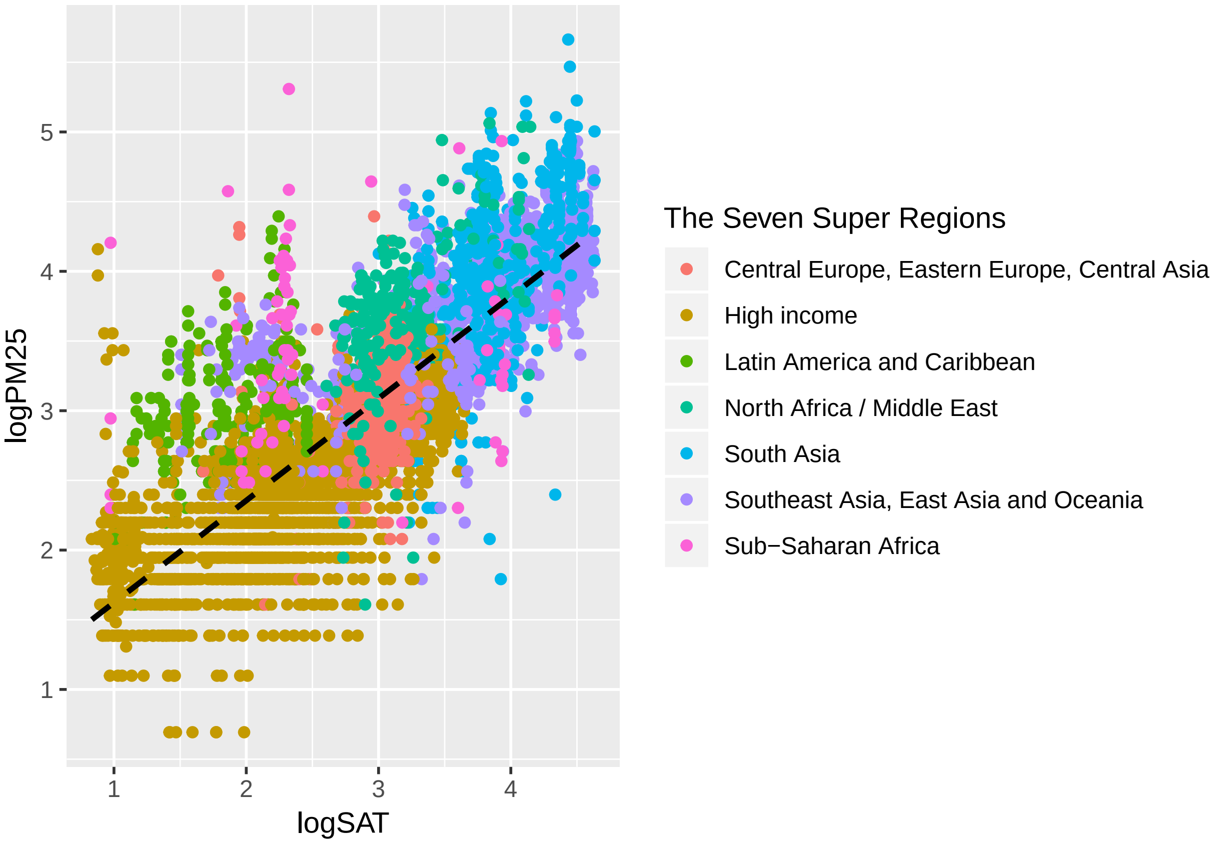
Where $P(\theta)$ in above equation is definded as a *Prior*, and $P(\theta|y)$ is a *posteriori*. In order to get the posteriori, some prior should be set first. Prior contains the initial belief of the parameter. It can be any distribution as long as it is not dramatically wrong.

3.2. Statistical Modelling

The single covariate is the logarithm of satellite data, $LogSAT_s$, and the dependent variable is the logarithm of ground measurement of PM 2.5, $LogPM25_s$, avaiable at a discrete set of N_S countries.

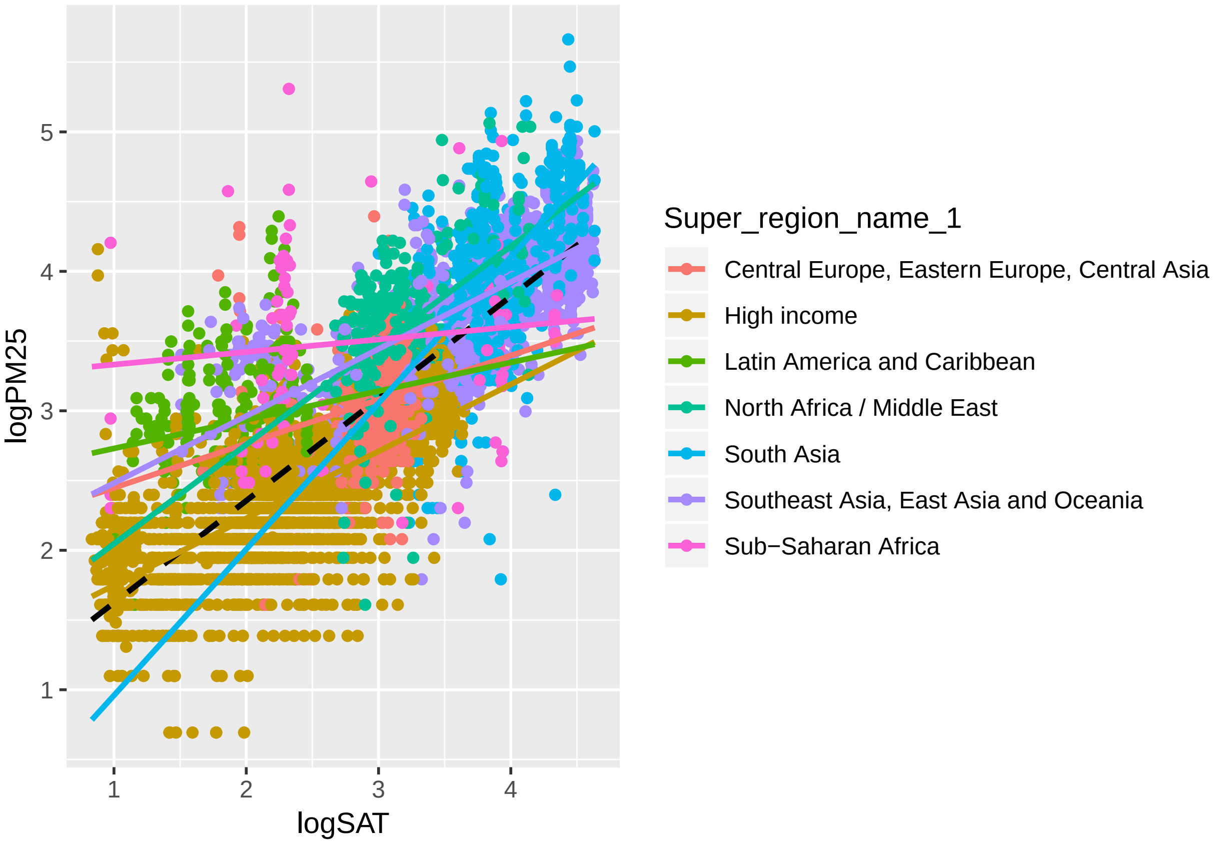
A naive yet stratforward way is the linear regression over all data in the dataset. For the simple linear regression, an intercept β_0 and a slope β_1 is estiamted using all the data(i.e. for the whole globe).

$$LogPM25 = \beta_0 + \beta_1 LogSAT$$



As can be see above, the dashes line produced is dominated by high income countries. This could be due to their higher income and thus, more sufficient data.

What about fitting regional lines?



When a line is fitted for each region, a lot of regions do no have enough infomation to get a good regression line since there are too few data. For several super regions, a tiny R^2 is observed:

```
##   super_region    R2
## 1            1 0.033
## 2            2 0.478
## 3            3 0.032
## 4            4 0.289
## 5            5 0.389
## 6            6 0.472
## 7            7 0.015
```

3.3 Borrowing Strength

A way to compromise is partial pooling, meanning that for each region,some information can be borrowed from the global pool of data to make a better estimation. The model then becomes

$$Y_s = \tilde{\beta}_{0s} + \tilde{\beta}_{1s} X_s + \epsilon_s$$

where X_s is the covariate, Y_s is the observed data at location s. $\epsilon_s \sim N(0, \sigma_\epsilon^2)$ is a ramdon error term. $\tilde{\beta}_{0s}$ and $\tilde{\beta}_{1s}$ varies for each location s, enables the intercept and slope to differ over space.

$$\tilde{\beta}_{0s} = \beta_0 + \beta_{0s}$$

$$\tilde{\beta}_{1s} = \beta_1 + \beta_{1s}$$

Where β_0 and β_1 are the means of slope and intercepts. β_{0s} and β_{1s} are spatial random effects representing how each region s varies from these means. For each country i, $i = 1, ..., N_{jk}$ in region j, $j = 1, ..., N_j$, in super region k, $k = 1, .., 7$, the model is :

$$LogPM25_{ijk} = \tilde{\beta}_{0,ijk} + \tilde{\beta}_{1,ijk} LogSAT_{ijk}$$

$$\tilde{\beta}_{0,ijk} = \beta_0 + \beta^R_{0,jk} + \beta^S_{0,k} R$$

Here, the random effects $\tilde{\beta}_{0,ijk}$ and $\tilde{\beta}_{1,ijk}$ have contributions from the country, the region and the super region. Same for the slope:

$$\tilde{\beta}_{1,ijk} = \beta_1 + \beta^R_{0,jk} + \beta^{SR}_{1,k}$$

To make it more clear, for each parameter β, let β^{SR}_k denote the coeffecient for super region k, β^{SR}_k is distributed with a mean of β and variance σ^2_{SR}, representing the variablity among different super regions:

$$\beta^{SR}_k \sim N(\beta, \sigma^2_{SR})$$

where $k = 1, ..., 7$.
Within each super region, there are several regions,j. Let β^R_{jk} denote the coeffcient for region j in super region k, with mean β^{SR}_k and variance $\sigma^2_{R,k}$, representing the variability within the super region k:

$$\beta^R_{jk} \sim N(\beta^{SR}_k, \sigma^2_{R,k})$$

where $j = 1, ..., N_j$, the number of countries in the region. Similarly, $\tilde{\beta}_{0,ijk} = \beta^C_{ijk}$ represents the coefficient for country i in region j, super region k with mean β^R_{jk} and variance $\sigma^2_{C,jk}$, the coefficient and variaability within region j in super region k respectively:

$$\beta^C_{ijk} \sim N(\beta^R_{jk}, \sigma^2_{C,jk})$$

where i is the number of countries in region j in super region k.
For above model, both country effects within regions and regional effects within super regions are independent. However, it is more helpful to consider neibering countries instead of all the countries in the region j. This is because information from direct neibours are more informative than distant countries. For instance, it makes more sense for Thailand to borrow information from China, while borrowing from Japan does not help much since they are very different, although they are all in Asia. Thus, instead of region, the neibours of the country *C=ijk* is used. So instead of above distribution of β^C_{ijk}:

$$\beta^C_i | \beta^C_{-i} \sim N(\beta^C_{-i}, \frac{\sigma^2}{N_{neibour}})$$

where for country i given its neibours, it is distributed with mean β^C_{-i} and variance $\frac{\sigma^2}{N_{neibour}}$, the mean and variance of the spatial random effects of its neibours. This model allows for a different calibration in each country. Country effects within regions are spatilly correlated and regional effects within super regions are independent.

3.3.1 Priors

R-INLA is used to fit the models. Before fittin, priors need examination. The default priors of R-INlA are used to give simulations for the countries:

$$\beta \sim N(0, 100)$$

$$\beta^{SR}_k \sim N(\beta, \sigma^2_{SR}),\ \sigma^2_{SR} \sim exp(100)$$

$$\beta^R_{jk} \sim N(\beta^{SR}_k, \sigma^2_{R,k}),\ \sigma^2_{R,k} \sim exp(100)$$

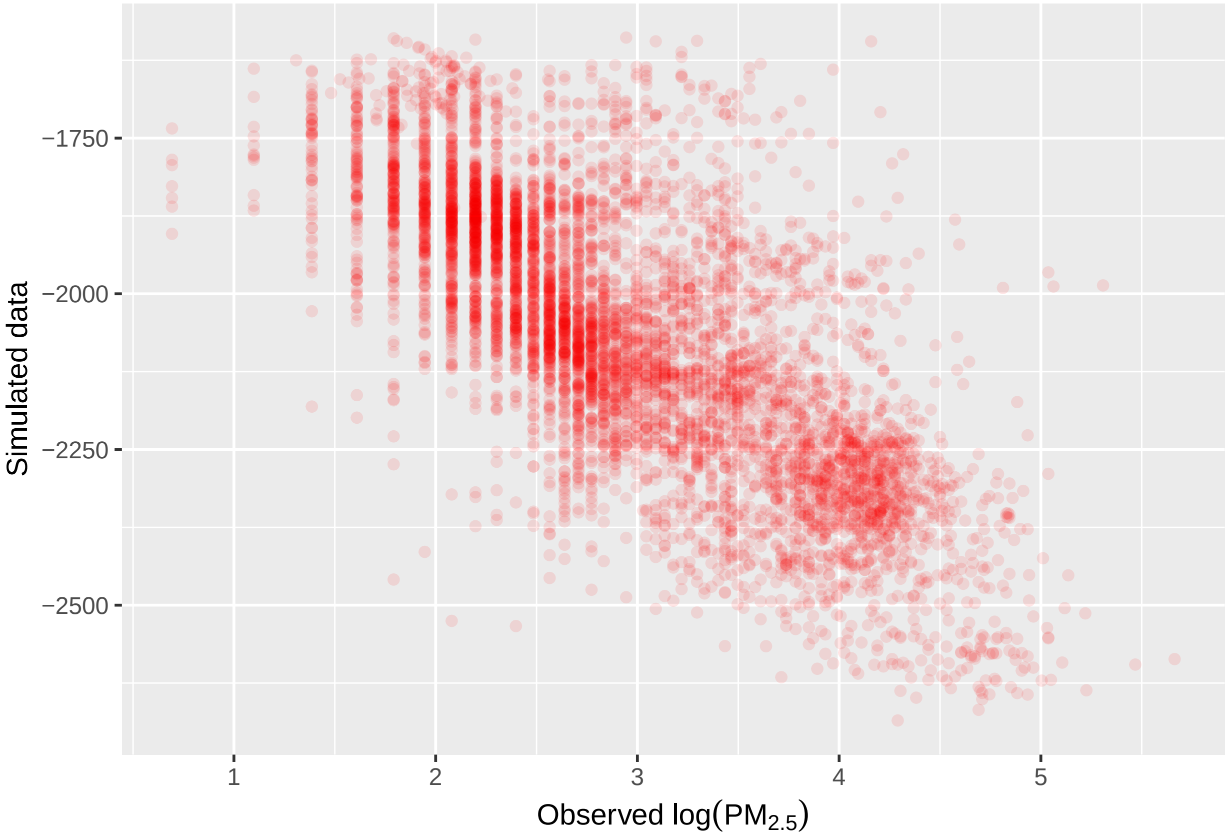
$$\mu \sim N(0, 1000)$$

$$\mu^{SR}_k \sim N(\mu, \sigma^2_{\mu,SR}),\ \sigma^2_{\mu,SR} \sim exp(100)$$

$$\mu^R_{jk} \sim N(\mu^{SR}_k, \sigma^2_{R,k}),\ \sigma^2_{R,k} \sim exp(100)$$

$$Y_{ijk} \sim N(\mu_{jk}^R + \beta_{jk}^R X_{ijk}, \sigma^2),\ \sigma^2 \sim exp(100)$$

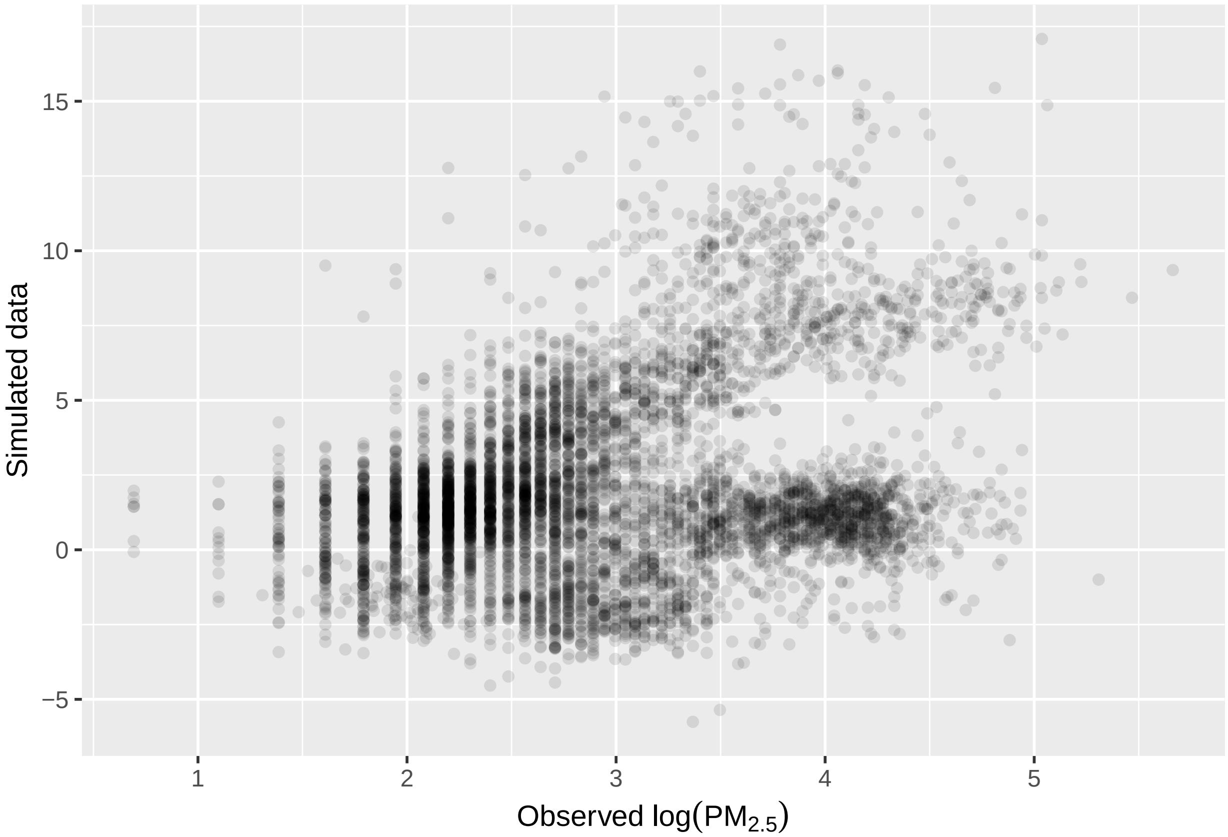
The log density of neutron star is only $60\mu gm^{-3}$. Such priors produces very unrealistic simulations:



On the other hand, with more sensible priors as following, the logPM2.5 simulated fall within [-20,20].

$$\beta \sim N(0,1)$$

$$\beta_k^{SR} \sim N(\beta, \sigma_{SR}^2),\ \sigma_{SR}^2 \sim N_+(0,1)$$
$$\beta_{jk}^R \sim N(\beta_k^{SR}, \sigma_{R,k}^2),\ \sigma_{R,k}^2 \sim N_+(0,1)$$
$$\mu \sim N(0,1)$$
$$\mu_k^{SR} \sim N(\mu, \sigma_{\mu,SR}^2),\ \sigma_{\mu,SR}^2 \sim N_+(0,1)$$
$$\mu_{jk}^R \sim N(\mu_k^{SR}, \sigma_{R,k}^2),\ \sigma_{R,k}^2 \sim N_+(0,1)$$
$$Y_{ijk} \sim N(\mu_{jk}^R + \beta_{jk}^R X_{ijk}, \sigma^2),\ \sigma^2 \sim N_+(0,1)$$



4. Models comparison and evaluation

Following from the discussin in section 3, three different models are fit for this data using R-INLA. The full model has spatially correlated country effects within neibours and independent regional effects within super regions. A *bym2* model is used to specify the spatial correlation between neibours.
The second model the same but both country effects within neibours and regional effects within super regions are independent. Since there is no regional data, this model is the same as using only super region as random effect. The third model is the simple linear regression model.

4.1 Leave-one-out Cross Validation

The goodness of the model is evaluated using conditional probability ordinate(CPO) and probability intergal transform(PIT).

4.1.1 Comparing CPO

CPO is a leave-one-out cross-validation score, showing the posterior probability of getting yi using the model fitted with all data except yi. For each observation:

$$CPO_i = p(y_i|y_{-i})$$

where y_{-i} is the set of obersevations ommiting the i^{th} obersevation. In INLA, the scoring rule of a model is

$$CPO = -\frac{1}{2}\sum_i log(p(y_i|y_{-i}))$$

The smaller CPO is, the better the model fits.

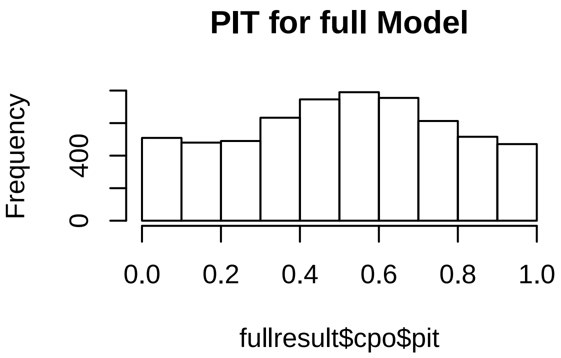
For the three models, $CPO_{full} = 815.4686902$, $CPO_{iid} = 1174.8636846$, $CPO_{slr} = 1955.9805031$ Thus the full model, with the smallest CPO score, is the model of best fit among the three. The simple linear model, with the lowest score, is not doing a good job.

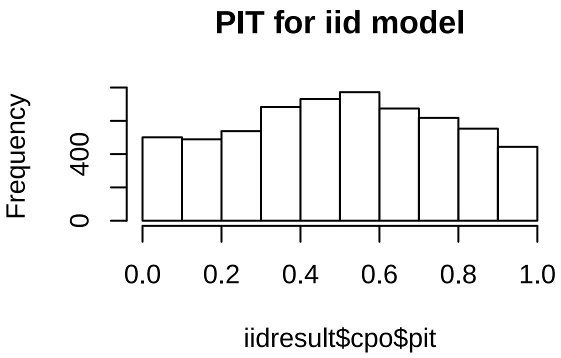
4.1.2 Compairing PIT

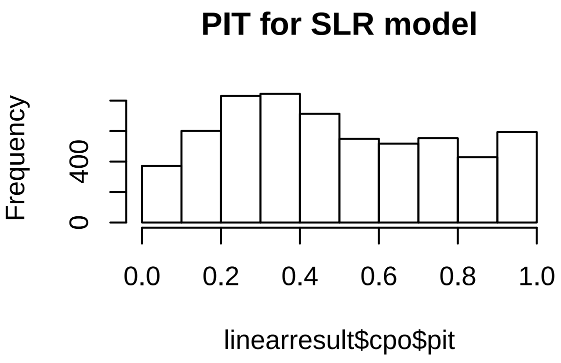
The cross-validated probability integral transform (PIT) is also a leave-one-out cross-validation score

$$PIT_i = p(y_i^* <= y_i | y_{-i})$$

where y_{-i} is the set of obersevations ommiting the i^{th} obersevation. For PIT, a uniform distribution is desired since by **the fundamental theorem of samping**, if $y_i \sim G_i$ and $G_i(y_i) \sim uniform(0, 1)$, the model is callibrated.


PIT for full Model


PIT for iid model


PIT for SLR model

It is noticed that the prediction intervals are too wide for all three models since the PIT graph concave up in the middle. The linear model is left skrewed a bit, showing that the posterior could be biased.

4.2 Cross Validation leaving more than one point

It has the same idea as leave-one-out CV to leave more data points out instead of just one. For the data that are not independent, Cross Validation that leaves more than one point is better since it only requires the blocks to be independent. Fot the PM 2.5 data, countries within each neighbours are dependent while the countries that are not neighbouring and super regions are independent. Most of the data comes from certain super regions. To evaluate the models, cross validation is performed with 3 sets of training set(around 90% of the data) and test set. The mean square error(MSE) is calculated using the test sets.

Super_region_name_1	n	super_region_code
Central Europe, Eastern Europe, Central Asia	518	1
High income	3051	2
Latin America and Caribbean	308	3
North Africa / Middle East	273	4
South Asia	464	5
Southeast Asia, East Asia and Oceania	1312	6
Sub-Saharan Africa	77	7

Considering the spatial natural of the problem, since the data among different super regions are independent, a sensible way is to leave out data such that it is difficult to borrow strength from neighbours. The way split a test set with around 600 observations out is as follows:
1. Randomly select two super region with less than 600 observations, since otherwise it is easy for countries in the super region to borrow strength.(select among super regions 1,3,4,5,7)
2. If the sum of the two super regions observations is less than 600, make it close to 600 while leaving several observations.
3. Among selected regions, randomly select observations to put to the test set. The goal is to leave few observation within the super regions. Since most observations are removed within one super region, the neighbouring data are also mostly removed, it is harder to borrow strength. The random sample produces training/test split as follows:
Training set 1: Randomly put 500 observations from SR1 and 100 observations from SR4 to test set.
Training set 2: Randomly put 270 observations from SR4 and 300 observations from SR3 to test set.
Training set 3: Randomly put 300 observations from SR5 and 300 observations from SR3 to test set. For each model, three training sets are used to build three models. For each fitted models, 1000 posterior parameter samples are drawn randomly. Those samples are then used on the test sets. An error is calculated by sum of $(mean(y_{estimate}) - y_{true})^2$ The resulting MSE is as follows:

model	MSE
full	0.273
iid	0.195
SLR	0.382

Unlike the CPO check, based on the MSE, the linearly independent model is the best in case of out of sample prediction. The SLR model is still the worst model of the three. This could be because when it is difficult to borrow strength from neighbours, the estimate based on regions are better.

Reference

Shaddick, G., M. L. Thomas, A. Green, M. Brauer, A. Donkelaar, R. Burnett, H. H. Chang, A. Cohen, R. V. Dingenen, C. Dora, et al. (2017). Data integration model for air quality: a hierarchical approach to the global estimation of exposures to ambient air pollution. Journal of the Royal Statistical Society: Series C (Applied Statistics) Available online 13 June 2017. arXiv:1609.00141.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., Gelman, A. (2018). Visualization in Bayesian workflow. Journal of the Royal Statistical Society Series A, accepted for publication. arXiv preprint: https://arxiv.org/abs/1709.01449