Bayesian Analysis Final Report

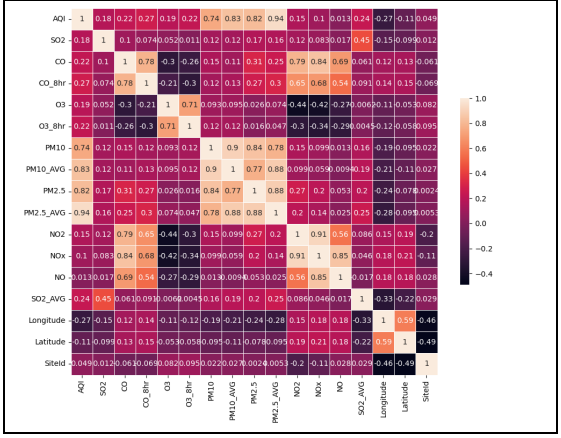
Group 10: 楊沛蓉、吳尚明、陳冠裕

1. Key Concepts

(A)特徵選定：

為了要預測 PM2.5，所以我們想知道什麼特徵和 PM2.5 較相關，透過相關的特徵值做預測。我們利用 2023 年 4 月份的資料做分析，並搭配 Heat map 查看 PM2.5 和其他特徵值的相關程度。透過 Heat map 可以看到 PM2.5 和 PM2.5_AVG、PM10 和 AQI 的相關程度最高，此三個特徵的相關度都大於 0.8。

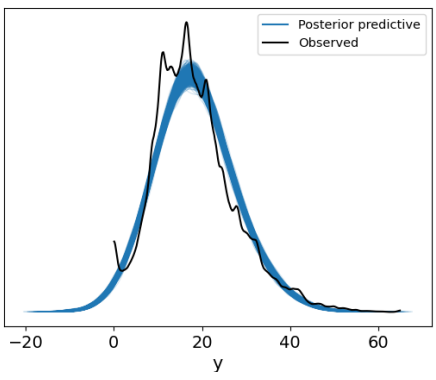
Heat Map



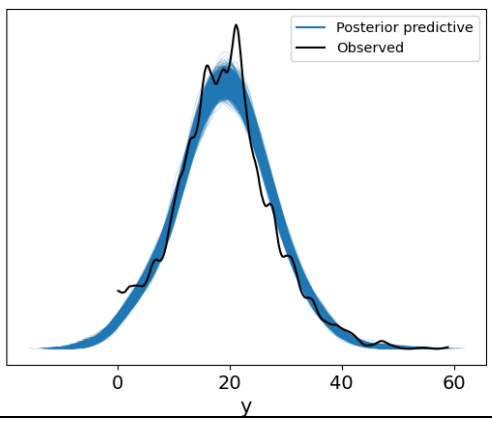
選定相關特徵後，我們利用 Mean Square Error 和 Posterior Predictive Check 決定訓練資料集的大小。評估此兩種方法的結果後，我們決定將訓練資料量設為 3 天。

Data Size	Training MSE	Testing MSE
7 days	25.9214	35.8212
3 days	14.3556	28.2673
1 day	65.9923	114.0947

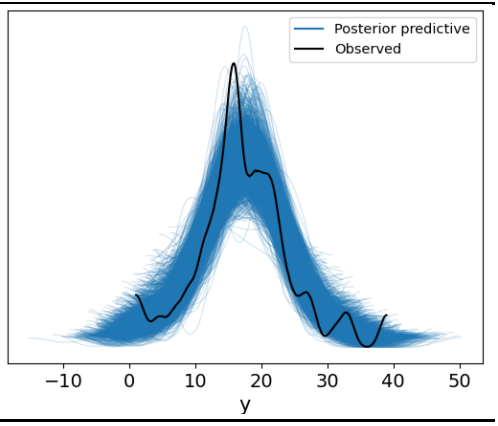
I. 訓練資料集大小為 7 天



II. 訓練資料集大小為 3 天



III. 訓練資料集大小為 1 天



(B) **資料前處理**

我們比較 3 種處理缺失資料的方法：1. Drop 2. Forwarding 3. Fill with mean value。第一種 Drop 的方法為直接捨棄含有缺失值的資料。第二種 Forwarding 的方法為複製在空白值前，且距離最近的非空白值進此空白裡。第三種 Fill with mean value 的方法是計算此特徵的平均值，並將此特徵的所有空白值利用此平均值填補。我們比較 3 種方法的標準化前的 MSE 和標準化後的 MSE 發現，標準化資料能使模型有較小的 MSE。

2. Performance of Models

我們利用了 Mean Square Error、R Square、Posterior Predictive Check 及 Expected Log Pointwise Density 比較 3 種模型 (Multi-Linear Regression Model、Polynomial Linear Regression Model、Hierarchical Linear Regression Model)。以下為使用 4 種比較方法的結果。利用 Mean Square Error 及 R Square 可以清楚看到 Multi-Linear Regression Model 有較好的表現。從 Posterior Predictive Check 的圖可以看到 Multi-Linear Regression Model 和 Hierarchical Linear Regression Model 沒有顯著差異，因為此兩種模型都是使用 Forwarding 作為資料前處理的方式，且都使用兩種 Feature(PM2.5_AVG, PM10)。另外，從 Expected Log Pointwise Density 中可以看到 Multi-Linear Regression Model 和 Hierarchical Linear Regression Model 的值相差很小，且比 Polynomial Linear Regression Model 的數值還大。越高的 Expected Log Pointwise Density 即表示模型有越好的預測表現，從此可以推斷 Multi-Linear Regression Model 和 Hierarchical Linear Regression Model 在預測方面比 Polynomial Linear Regression Model 的表現還要好。

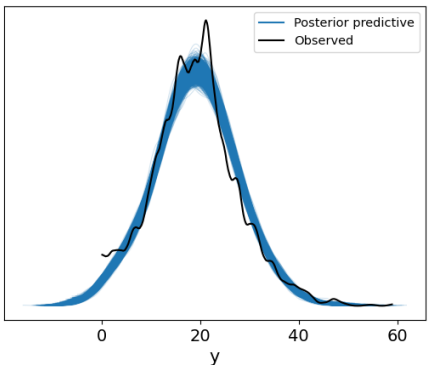
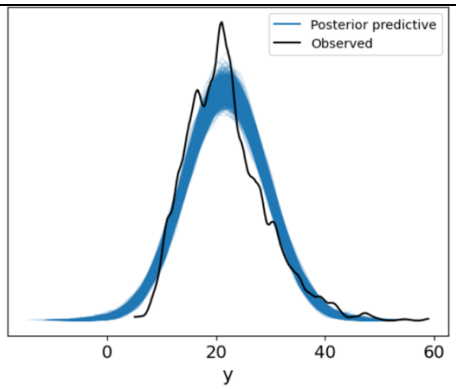
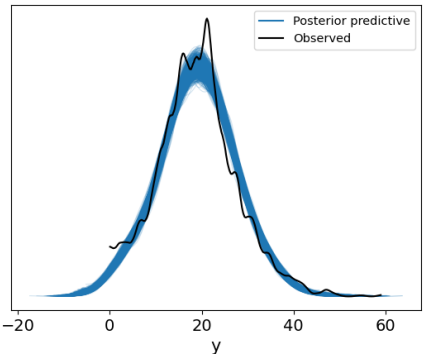
(A) Mean Square Error

Model	Training MSE	Testing MSE
Multi-Linear Regression	14.3556	28.2673
Polynomial Linear Regression	44.0998	51.6605
Hierarchical Linear Regression	93.0863	116.4665

(B) R Square

Model	Training R Square	Testing R Square
Multi-Linear Regression	0.8051	0.781
Polynomial Linear Regression	0.3563	0.546
Hierarchical Linear Regression	0.5064	0.6686

(C) Posterior Predictive Check

Multi-Linear Regression Model	Polynomial Linear Regression Model	Hierarchical Linear Regression Model
		

(D) Expected Log Pointwise Density

Multi-Linear Regression Model	Polynomial Linear Regression Model	Hierarchical Linear Regression Model

```
            Estimate       SE
deviance_loo 20757.72   115.47
p_loo            5.65        -
------

Pareto k diagnostic values:
                         Count   Pct.
(-Inf, 0.5]   (good)     3773  100.0%
 (0.5, 0.7]   (ok)          0    0.0%
   (0.7, 1]   (bad)         0    0.0%
   (1, Inf)   (very bad)    0    0.0%
```

```
            Estimate       SE
deviance_loo 18576.82    86.77
p_loo            5.53        -
------

Pareto k diagnostic values:
                         Count   Pct.
(-Inf, 0.5]   (good)     2838  100.0%
 (0.5, 0.7]   (ok)          0    0.0%
   (0.7, 1]   (bad)         0    0.0%
   (1, Inf)   (very bad)    0    0.0%
```

```
Computed from 4000 posterior samples and 3773 observati

            Estimate       SE
deviance_loo 20757.36   115.41
p_loo            5.44        -

There has been a warning during the calculation. Please
------

Pareto k diagnostic values:
                         Count   Pct.
(-Inf, 0.5]   (good)     3770   99.9%
 (0.5, 0.7]   (ok)          0    0.0%
   (0.7, 1]   (bad)         3    0.1%
   (1, Inf)   (very bad)    0    0.0%
```

3. Conclusion

透過比較模型，可以得出結論：利用 Multi-Linear Regression Model，使用 2 個特徵(PM2.5_AVG、PM10)，並搭配 Forwarding 方法做資料前處理，再進行資料標準化，會有比 Polynomial Linear Regression Model 及 Hierarchical Linear Regression Model 較佳的表現。