貝氏資料分析Final Project

作業環境：Pymc4

資料集選擇：2021.01～2022.10

資料前處理：

1.將將publishtime轉換成和2021/1/1對比的天數，以方便model訓練時做判斷

ex: 2021/1/10就會被轉換成9

2.將一天的各項數據轉換成平均，減少數據集大小

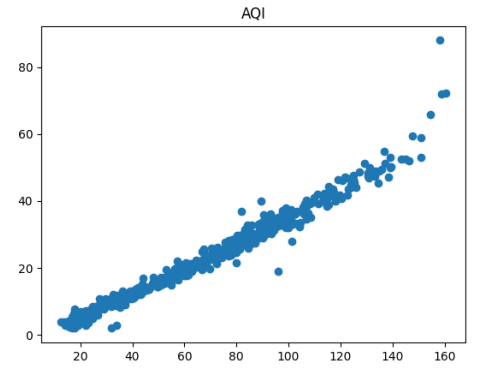
ex: 2021/1/1的PM2.5濃度為整天平均

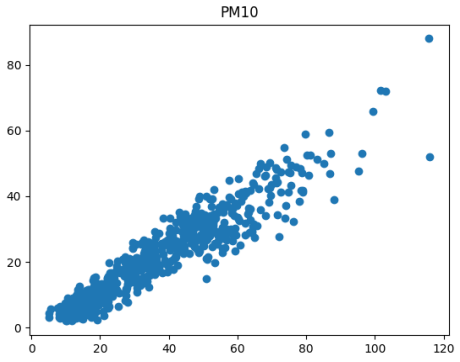
3.由於政府資料的各站點的site ID有缺失，例如：總共只有65個測站但有66個代碼，因此我們對各個測站做encoding，共92個不同測站，將各個測站轉換成代碼

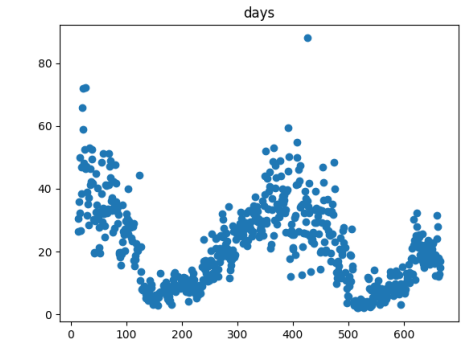
ex: 鳳山編號為89，三義為編號0

feature選擇：

SO2, NO2, AQI, PM10, time, ID, Days作為模型feature，選擇的原因如下圖，AQI、PM10及days都與PM2.5有著線性關係或sin函數關係（SO2及NO2的選擇亦同），time、ID及Days則是因為地區與時間本身就會對PM2.5造成影響，因此選擇。





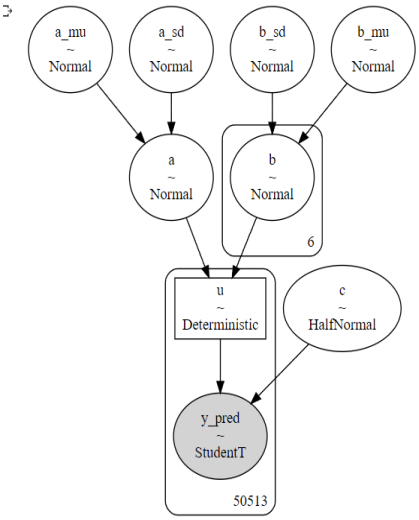


模型（下圖是我們的mutiple linear regression）：

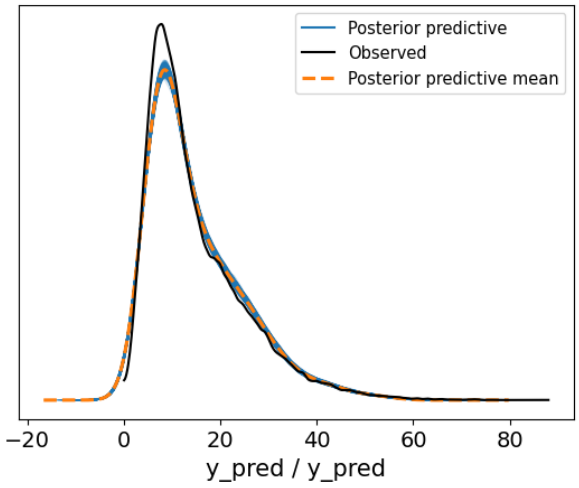
1.大部分都是呈現線性關係，唯獨時間對PM2.5呈現sin函數分布，因此在Deterministic部分，days係數前使用了sin function
2.使用Hierarchical Model
3.偶爾會有outlier的出現所以使用StudentT去處理

```
with pm.Model() as final_model:
    a_mu = pm.Normal('a_mu', mu=0, sigma=10)
    a_sd = pm.Normal('a_sd', mu=0, sigma=10)
    a = pm.Normal('a', mu=a_mu, sigma=a_sd)
    b_mu = pm.Normal('b_mu', mu=0, sigma=10)
    b_sd = pm.Normal('b_sd', mu=0, sigma=10)
    b = pm.Normal('b', mu=b_mu, sigma=b_sd, shape=6)
    c = pm.HalfNormal('c', sigma=10)
    u = pm.Deterministic('u', a+b[0]*np.sin(days)+b[1]*so2+b[2]*no2+b[3]*id+
                                        b[4]*aqi+b[5]*pm10)
    y_pred = pm.StudentT('y_pred', mu=u, sigma=c, nu=15, observed=train_y)
    trace = pmjax.sample_numpyro_nuts(500, tune=2000, target_accept=0.8)
    ppc = pm.sample_posterior_predictive(trace)
```

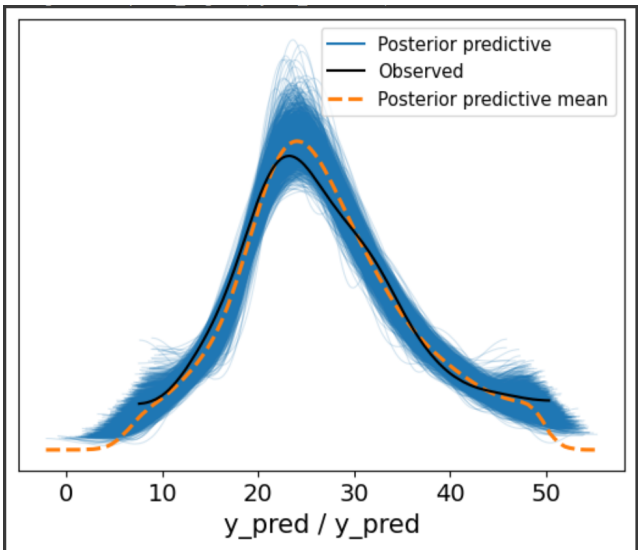
Kruschke:



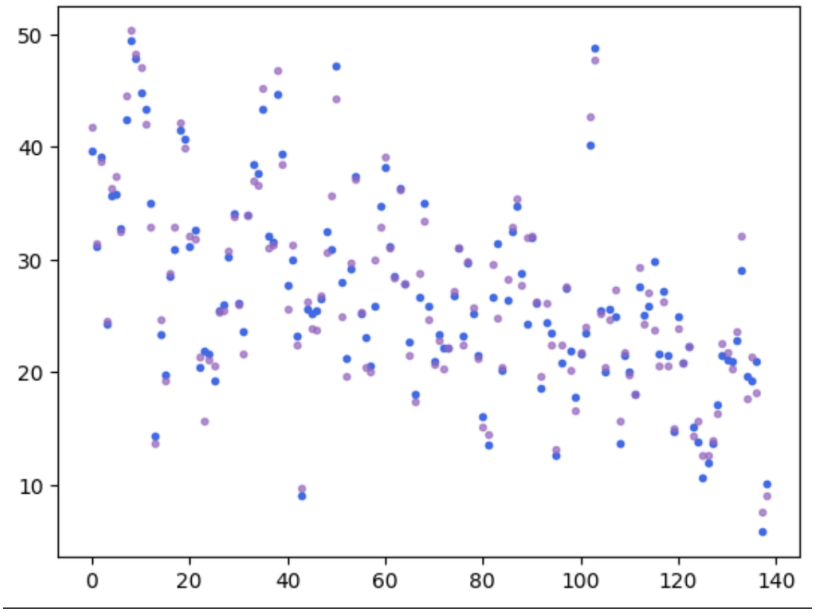
Posterior predict check(within sample):



Posterior predict check(Out of sample):
資料選用2023.1～2023.5 鳳山



Predict Vitualization:
紫色點是實際PM2.5的位置
藍色點是我們預測的PM2.5
可以看到基本上大多數的趨勢都有掌握住，預測結果相當理想。



工作分配:
林宥騰:資料前處理，視覺化
陳致齊:model
邱子恩:上台報告、報告撰寫