Final Project

HeJin Chu STAT-618-001

Multilevel Models in Quantitative Research

1. Description of the Dataset:

   The dataset I used for this project is from Kaggle. This data set is about the weekly sales amount of 45 different store of Walmart from 2010 to 2012. There are 6,435 observations in this dataset. After deleting missing data, it has 2,565 observations.

In this dataset, there are 8 variables as following:

1. Store: Store ID
2. Date: Date of the sales
3. Weekly_Sales: Amount of weekly sales
4. Holiday_Flag: Whether the week is a holiday week
5. Temperature: Average temperature of the week
6. Fuel_Price: Cost of the fuel in the region of the week
7. CPI: Consumer Price Index of the week
8. Unemployment: Unemployment rate of the week

- Missing Data:

There are some missing data appeared after I used as.Date function to make variable "Date" proper to be analyzed. Because the missing data didn't exist in the original data, I decided to delete missing value directly.

- Source of the data:

https://www.kaggle.com/code/ariosliew92/walmart-sales-analysis-multilevel-modelling

2. Research Topic:

My research question is "What is the best varying-intercept & varying- slope model to explain weekly sales amounts for theses 45 different store?". I want to find out which variables will exist in the best varying-intercept model.

3. Statistical model:

By building many different multilevel models, I can compare all models and find out what the best model I want. After comparing, I can analyze the best model to realize its detail, like the coefficient for each variable and other important information.

First, I build many varying-intercept models, including simple model and model with all variables.

Its formal equation is:

$$Y_{ij} = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \cdots + u_{oj} + \epsilon_{ij}$$

Where

Y_{ij} is the i-th sales amount in the j-th store

β_0 is intercept for fixed effect

β_1 is the coefficient of X_{1ij}, which means the slope of the first independent variable for fixed effect

β_2 is the coefficient of X_{2ij}, which means the slope of the second independent variable for fixed effect

u_{oj} is the random effect of the j-th store

ϵ_{ij} is residual

Then, I add different varying-slope into the best varying-intercept model.

Its formal equation is:

$$Y_{ij} = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \cdots + u_{oj} + u_{1j} X_{1ij} + u_{2j} X_{2ij} + \cdots + \epsilon_{ij}$$

Where

Y_{ij} is the i-th sales amount in the j-th store

β_0 is intercept for fixed effect

β_1 is the coefficient of X_{1ij}, which means the slope of the first independent variable for fixed effect

β_2 is the coefficient of X_{2ij}, which means the slope of the second independent variable for fixed effect

u_{oj} is the intercept for random effect of the j-th store

u_{1j} is the coefficient of X_{1ij}, which means the slope of the first independent variable for random effect

u_{2j} is the coefficient of X_{2ij}, which means the slope of the second independent variable for random effect

ϵ_{ij} is residual

4. Model Diagnostics:

After building all varying-intercept models, I make a table to compare the AIC for each model:

Model <chr>	AIC <dbl>
Sale.Store	66788.51
Sale.Date	66783.87
Sale.Holiday	66771.05
Sale.Temp.Fuel	66761.94
Sale.CPI.Unemployment	66720.07
Sale.Store.full	66636.11

According to this table, we can know the final model, which is the model with all independent variables has the best performance to explain weekly sales amount since it has the smallest AIC.

Then, I also compare all varying-slope models after I add them into the best varying-intercept model.

Model <chr>	AIC <dbl>
Date.Store.full	66674.14
Holiday.Store.full	66726.91
Temp.Fuel.Store.full	66687.95
All.full	66302.50

4 rows

According to this table, we can know the varying-slope model with all variables has the best performance.

5. Empirical Finding:

Here is the information for the best varying-intercept model:

```
lmer(formula = Weekly_Sales ~ Date + Holiday_Flag + Temperature +
    Fuel_Price + CPI + Unemployment + (1 | Store), data = walmart)
              coef.est   coef.se
(Intercept)   1639734.96  232943.22
Date              -64.31      14.87
Holiday_Flag1    8644.13    6907.90
Temperature      -190.13     129.32
Fuel_Price     -10055.42    8083.73
CPI              4042.69    1120.21
Unemployment   -31933.06    4500.67

Error terms:
 Groups   Name        Std.Dev.
 Store    (Intercept) 592969.82
 Residual             101155.21
---
number of obs: 2565, groups: Store, 45
AIC = 66636.1, DIC = 66848.9
deviance = 66733.5
```

According to this result, we can know the formal equation of the best model is

$$Weekly\_Sales_{ij} = 1{,}639{,}734.96 - 64.31Date_{ij} + 8{,}644.13Holiday\_Flag_{ij} - 190.13Temperature_{ij} - 10{,}055.42Fuel\_Price_{ij} + 4{,}042.69CPI_{ij} - 31{,}933.06Unemployment_{ij} + u_{oj} + \epsilon_{ij}$$

According to this model, we can know the influence of each variable on weekly sales amount is different. Some variables impact hugely, such as fuel price, unemployment; other variables may have smaller effect, like date and temperature.

Then is the information for the best varying-intercept & slope model:

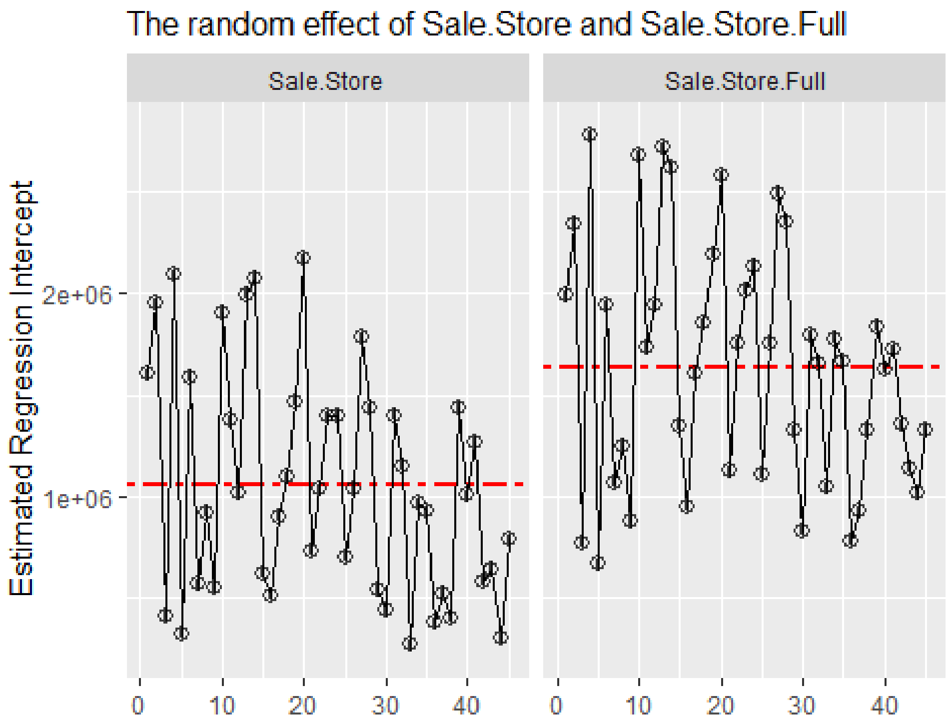
```
Linear mixed model fit by REML ['lmerMod']
Formula: Weekly_Sales ~ Date + Holiday_Flag + Temperature + Fuel_Price +
    CPI + Unemployment + (1 + Date + Holiday_Flag + Temperature +
    Fuel_Price + CPI + Unemployment | Store)
   Data: walmart

REML criterion at convergence: 66230.5

Scaled residuals:
    Min      1Q  Median      3Q     Max
-6.2907 -0.4487 -0.0407  0.3649  5.9794

Random effects:
 Groups   Name         Variance  Std.Dev.  Corr
 Store    (Intercept)  8.343e+09  91338.85
          Date         1.409e+02     11.87 -0.70
          Holiday_Flag1 5.637e+07   7508.14 -0.26  0.16
          Temperature  3.484e+06   1866.48  0.15  0.15 -0.63
          Fuel_Price   3.548e+08  18836.72 -0.34  0.49  0.84 -0.60
          CPI          2.742e+07   5236.45 -0.11 -0.08  0.65 -0.55  0.63
          Unemployment 1.323e+10 115008.39  0.10 -0.07 -0.05  0.04 -0.16 -0.42
 Residual              7.888e+09  88813.39
Number of obs: 2565, groups:  Store, 45

Fixed effects:
                Estimate Std. Error t value
(Intercept)   1464493.98  254523.51   5.754
Date              -76.08      14.01  -5.430
Holiday_Flag1    9722.40    6198.70   1.568
Temperature      -366.98     301.38  -1.218
Fuel_Price       4796.27    8148.00   0.589
CPI              4603.11    1694.40   2.717
Unemployment    -4730.49   18474.01  -0.256

Correlation of Fixed Effects:
            (Intr) Date   Hld_F1 Tmprtr Fl_Prc CPI
Date        -0.376
Holidy_Flg1  0.146 -0.317
Temperature -0.049  0.120 -0.108
Fuel_Price   0.508 -0.339  0.218 -0.282
CPI         -0.546 -0.422  0.158 -0.266 -0.173
Unemploymnt -0.269  0.062 -0.033  0.046 -0.070 -0.051
optimizer (nloptwrap) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')
```

According to this result, we can know the formal equation of the best varying-intercept & slope model is

$$
\begin{aligned}
Weekly\_Sales_{ij} &= 1{,}464{,}493.98 - 76.08Date_{ij} + 9{,}722.40Holiday\_Flag_{ij} \\
&- 366.98Temperature_{ij} + 4{,}796.27Fuel\_Price_{ij} \\
&+ 4{,}603.11CPI_{ij} - 4{,}730.49Unemployment_{ij} + u_{oj} \\
&+ +u_{1j}Date_{ij} + u_{2j}Holiday\_Flag_{ij} + u_{3j}Temperature_{ij} \\
&+ u_{4j}Fuel\_Price_{ij} + u_{5j}CPI_{ij} + u_{6j}Unemployment_{ij} + \epsilon_{ij}
\end{aligned}
$$

6. Figure:

This is the line plot comparison for the simplest model and the best model



According to this plot, we can see the random intercept estimates and confidence intervals for different stores show considerable variability, indicating significant inter-store variability. Moreover, the model on the right seems to have more concentrated confidence intervals for the stores compared to the model on the left, suggesting that the right-side model includes more variables, thereby explaining part of the variability between stores.

7. Conclusion:

According to above results, we can know both the best varying-intercept model and the best varying-intercept & slope model has all variables. Although each variable has different fixed or random effect for the weekly sales amount, all variables are necessary. If the model lacks any predictor, its explanatory power will decrease.