20-2 DATA SCIENCE(1) 파이널 프로젝트

쓰레기, 누가 더 많이 버릴까?

서울시 자치구별 음식물 쓰레기 배출량에 대한 Linear Mixed Effects 모델을 활용한 회귀분석, 예측 및 정책제언

11 DEC 2020

2013121170 강경훈

Table of Contents

I. 데이터와 Research Question

음식물 쓰레기 갈수록 많아진다. 심각하다!

Research Question: 자치구별 월평균 음식물 쓰레기 배출량 변동을 설명하고 예측해보자.

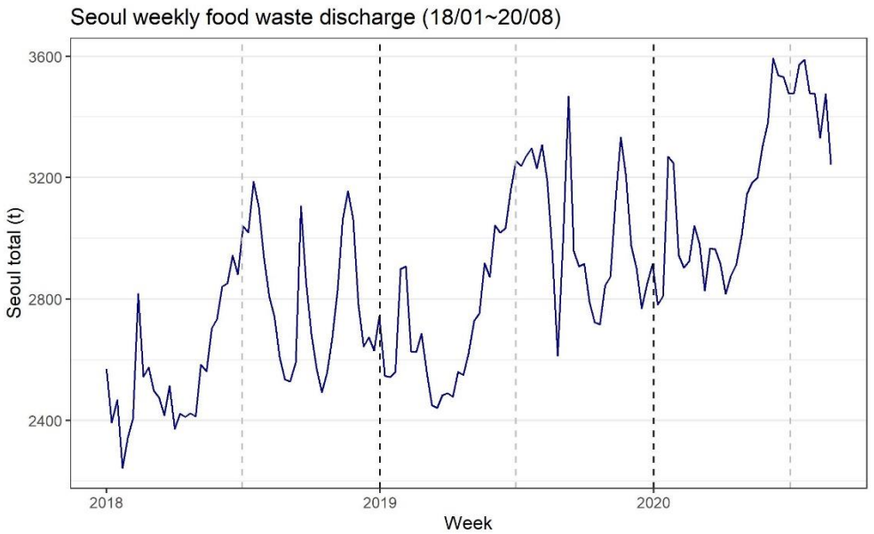
1. Data

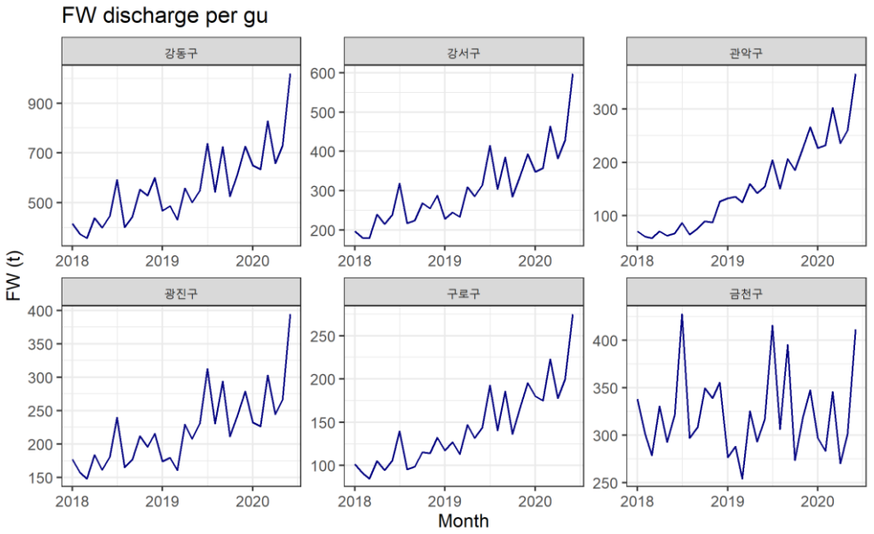
Schema 1 (690 X 14)		Y(i,j)	X1(i,j)	X2(i,j)	X3(i,j)	X4(i,j)	X5(i,j)	X6(i,j)	X7(i,j)	X8(i,j)	X9(i,j)	X10(i,j)	X11(i,j)	X12(i,j)	X13(i,j)
Group(i)	Index(j)	배출량	세대수	평균세대원	18세이하	65세이상	1인비율	점포수	점포생존율	소득수준	NO2	pm10	평균기온	배달검색	명절

강동구	2018년 1월	인구 (수요)	세대수	지역 내 총 세대 수	식당 (공급)	점포수	지역 내 총 점포 수
	2018년 2월		평균세대원	지역 내 세대별 평균 가구원 수		점포생존율	지역 내 신생 점포 중 영업기간 1년 이상 비율
	…		18세이하	지역 내 인구 중 18세 이하 비율		소득수준	지역 내 평균 소득 수준 (분위)
	2020년 6월		65세이상	지역 내 인구 중 65세 이상 비율			
강서구	2018년 1월		1인비율	지역내 세대 중 1인가구 비율			
	2018년 2월				환경	NO2	지역 내 일평균 이산화질소 측정치
	…	기타	배달검색	"배달" 검색 상대빈도 (기간 내 최대치 100)		pm10	지역 내 일평균 미세먼지 측정치
	2020년 6월		명절	기간 내 추석, 설날 포함 여부		평균기온	지역 내 일평균 기온
…	…	날짜를 Integer로 바꾸어 설명변수로 사용					

- 자치구별 (강남, 강북, 용산 제외) 18년 1월 ~ 20년 6월 월평균 RFID 음식물 쓰레기 배출량 및 관련 변수

1. EDA





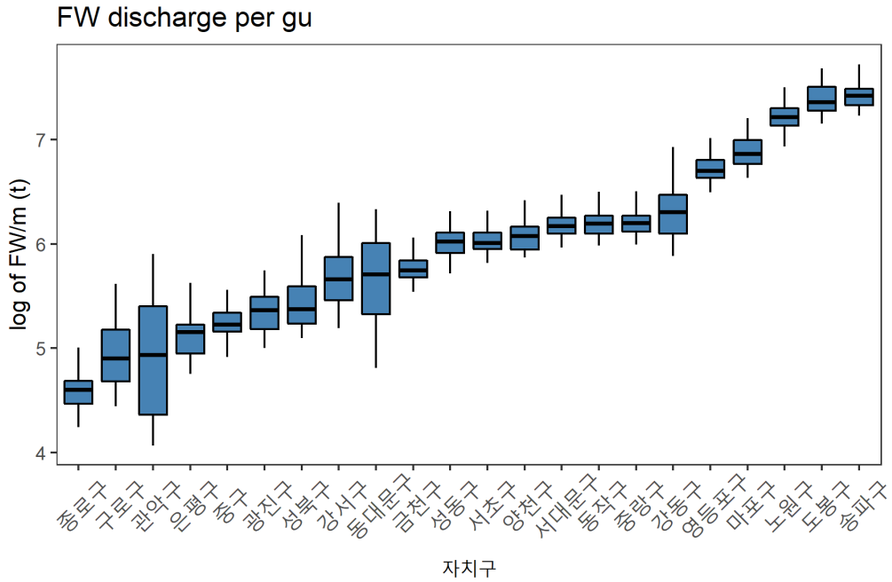
- 서울시 전체는 증가, 그러나 자치구별로 증가 속도와 수준이 다르다.
- **자치구별 배출량 관련 변수의 분포와 배출량과 분포의 관계가 다를것?!**

II. 방법론: Linear Mixed Effects Model

Complete pooling과 separate model (No pooling) 사이의 Partial pooling!

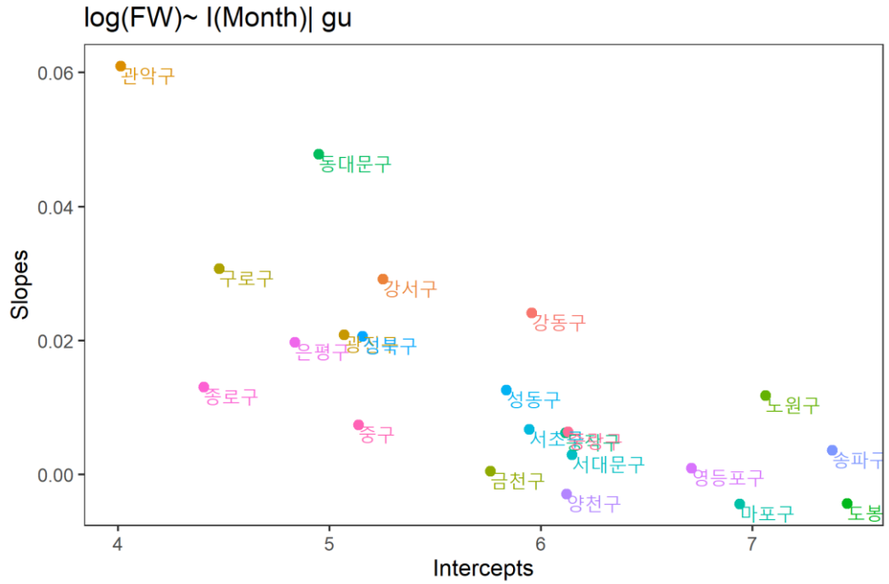
Response feature analysis 결과 Partial pooling이 적합한 것으로 생각됨, 널리 사용되는 방법 중 하나인 LMM 채택

- **Complete pooling?** 자치구별 차이 (group effect)를 고려하지 않고 모든 관측치에 대해 분산의 iid 가정, 하나의 회귀분석



- 자치구별로 평균이 판이하게 다르다. **즉 group effect가 명백히 존재하며, 개별 그룹 내 관측치들은 서로 correlated**

- **No pooling?** 각 자치구별로 분산의 분포가 완전히 다르다고 가정하여 자치구별로 회귀분석



- **자치구별 회귀계수에 상관관계 존재.** 계수들 간의 Exchangeability 가정하면, 어떤 공통된 분포의 iid 샘플로 볼 수 있음

- **Partial Pooling: 하나의 회귀 모델을 세우되, 그룹 내의 분산과 그룹 간의 분산을 모두 반영해야 함!**
  - **Bayesian:** 그룹 내 exchangeability, 그룹별 모수 간 exchageability 를 모두 prior로 모델링하는 Hierarchical Linear Regression Model
  - **Frequentist:** 회귀계수를 Fixed effect (모수), Random effect (확률변수)로 나누어 분산을 모델링하는 Linear Mixed Effects Model

When Mixed Effects (Hierarchical) Models Fail: Pooling and Uncertainty | by Eduardo Coronado Sroka | Towards Data Science

II. 방법론: Linear Mixed Effects Model

Linear regression의 conditional mean의 모델링이면, LME는 분산을 모델링!

종속변수에 대한 설명변수의 영향은 고정 효과 + (그룹별로 다른) 랜덤 효과, 때문에 그룹별로 다른 회귀계수를 "예측"할 수 있음

- **LMM: 선형식과 분포 가정 (i는 그룹 index)**

$$\boldsymbol{Y_i = X_i\beta + Z_i\gamma_i + \epsilon_i} \quad \boldsymbol{(\epsilon_i \sim N(0, \sigma^2 I_{n_i}) \perp \gamma_i \sim N(0, D))}$$

Diagnostics: 모델 학습 후 Residual과 random effect estimates가 이 조건을 만족하는지 확인해야 함 (e.g. qqnorm)

  - *Ex)* 개별 관측치 y_{ij} $(i \in [N], j \in [n_i])$와 설명변수 $x_{ij}^{(1)}, x_{ij}^{(2)}, x_{ij}^{(3)}$의 선형 관계식은 다음과 같이 쓸 수 있음

$$y_{ij} = \beta_0 + \gamma_{i0} + \beta_1 x_{ij}^{(1)} + (\beta_2 + \gamma_{i1}) x_{ij}^{(2)} + \gamma_{i2} x_{ij}^{(3)} + \sigma^2$$

    - β는 "추정"하는 모수(상수)로, **모든 그룹에 걸쳐 같음 (fixed effects)**
    - γ_i는 "예측"하는 확률변수로, **그룹에 따라 다름 (random effects)**
    - One-way ANOVA는 이론적으로 LMM $y_{ij} = \beta_0 + \gamma_{i0} + \sigma^2$와 같음

- **γ_i가 그룹별로 다르기 때문에 각 그룹마다 다른 선형식을 맞출 수 있음. 그룹 별 Prediction이 가능하다!**

$$\boldsymbol{Y_i \mid \gamma_i \sim N(X_i\beta + Z_i\gamma_i, \sigma^2 I_{n_i})}$$

(γ_i는 확률변수이므로, 모수 $\theta = \{\beta, D, \sigma^2\}$ 추정 이후 그룹별 조건부 분포 $\gamma_i \mid Y_i, \hat{\theta}$ 의 conditional mode로 예측)

- Marginal Y_i의 분포를 보면 LMM는 **종속변수의 분산에 group effect를 반영**$(V_i = \sigma^2 I_{n_i} + Z_i D Z_i^T)$**하는 방법임을 알 수 있음.**

$$\boldsymbol{Y_i \sim N(X_i\beta, \sigma^2 I_{n_i} + Z_i D Z_i^T)}$$

(Extending the Linear Model with R, CH10~11, Faraway, 2016) https://englianhu.files.wordpress.com/2016/01/faraway-extending-the-linear-model-with-r-e28093-2006.pdf

II. 방법론: Linear Mixed Effects Model

추정과 변수 선택이 어렵긴 한데, 데이터 적당히 많고, 컴퓨터 좋고 시간 많으면 괜찮다!

그룹 별 prediction이 가능하다는 장점이 있으나, 모수와 신뢰구간 추정, 검정, 변수 선택이 난해한 것이 단점

- **추정**: Newton이나 EM 등 알고리즘으로 MLE 구할 수 있으나, $l(\beta, D, \sigma^2 \mid \{Y_i\}) = \sum_i^N \left\{ -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|V_i| - \frac{1}{2}(Y_i - X_i\beta)^T V_i^{-1}(Y_i - X_i\beta) \right\}$
  1. **Boundary fit:** 분산 모수가 0보다 커야하는 boundary condition 때문에 MLE에서 로그 우도의 기울기가 0이 아닐 수 있다. 때문에 기울기 보고 찾는 optimization 방법 시행에 문제가 있을 가능성
     - 같은 이유로 MLE에서의 Hessian으로 SD를 추정하는 방법도 문제가 있음. 때문에 특히 Random effect 추정치의 신뢰구간 도출이 어려움 ***Nonparametric Bootstrap CI (data resampling) 으로 우회***
  2. **Biasedness:** MLE 자체가 분산을 과소 추정하는 경향이 있음, 그룹의 수가 적으면 이러한 편향이 심화 ***종속변수의 선형변환*** $(K^T Y_i \;\; s.t. \;\; K^T X_i = 0)$ ***으로 먼저 RE 모수를 추정하고 나중에 FE를 추정하는 REML로 개선 가능***

- **검정:** 분산 모수에 대한 가설 검정 시 검정통계량의 Null approx. dist의 수렴 조건에 위배
  1. **LRT for model comp:** $2(l(\hat{\boldsymbol{\theta}} \mid Y) - l(\theta_0 \mid Y))$에서, 분산 모수의 경우 주로 귀무가설이 모수공간의 boundary에 있기 때문에 $(H_0: \sigma^2 = 0)$ LRT의 Null distribution의 극한 분포$(\sim \chi^2)$를 쓰기에는 문제가 있음. ***LRT의 parameteric Bootstrap 분포로 우회***
  2. **F-test for fixed effects:** RSS와 df를 이용해 F-test를 할 때, random effect 때문에 df가 명확하지 않으며, 검정통계량이 F 분포라는 보장이 없음. ***Kenward-Roger adjusted F-test로 개선***

- **변수 선택:** 여러 모델의 비교 시 Multiple testing 문제 때문에 검정은 부적합. $\boldsymbol{AIC = -2Max\, l(\theta \mid y) + 2p}$ ***사용***
  1. 모수 개수 p에 RE를 어떻게 반영할 것인지 불분명함
  2. 일반적으로 Likelihood를 비교하는 변수 선택 방법은 Boundary condition이 있을 경우 신뢰성이 떨어짐.

그래도 어쩔 수 없다 AIC 써야지... 다른 방법이 딱히 없음

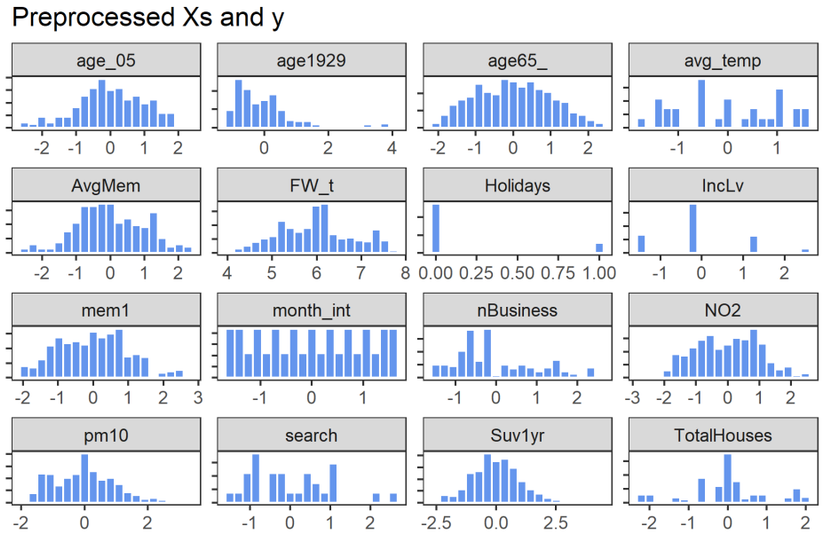
(Extending the Linear Model with R, CH10~11, Faraway, 2016) https://englianhu.files.wordpress.com/2016/01/faraway-extending-the-linear-model-with-r-e28093-2006.pdf

III. 데이터 전처리

모델 학습 전에 챙겨야 하는 준비물

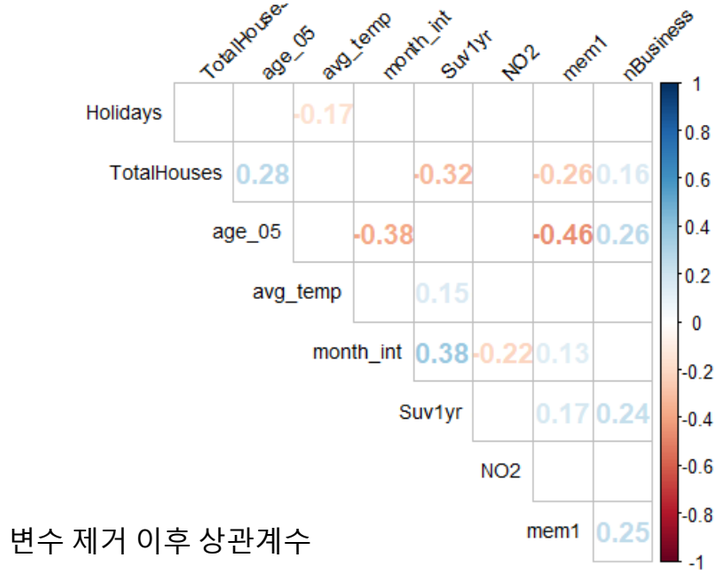
Pre-processing: 알고리즘의 Numerical 안정성을 위해 변수를 standardize, 모델 추정치의 안정성을 위해 일부 변수 제거

Data Logarized, Centered & Rescaled



- 종속변수 로그변환: marginal 분포가 정규분포가 되는 효과 외에도, 음식물 쓰레기의 변화량을 모델링 ($\log \frac{x_{t+1}}{x_t} \approx \frac{\Delta x_t}{x_t}$) 하는 의미가 있다.

Dimension Reduction



변수 제거 이후 상관계수

- 상관계수가 높은 ($|\rho| > 0.5$) 변수 제거
  (AvgMem, age1929, pm10, IncLv, age65_)

Remark: 왜 차원축소 (PCA or FA)를 하지 않고 변수를 제거해버렸나?

- 차원축소는 이론적으로는 상관관계가 높은 변수를 제거하지 않고 hidden factor들로 압축한다는 점에서 데이터의 변동을 살리면서 변수를 압축할 수 있는 좋은 방법이나, **문제는 Communication이 어렵다는 것!** ***(PC가 뭐에요? "인구 팩터"가 무슨 말이에요? 내 변수 돌려줘요!)***
- 음식물 쓰레기 데이터는 1) 줄여야 하는 변수가 그다지 많지 않고 2) 변수들 상관관계 대한 어느정도 domain knowledge가 있음 (1인가구 비율이 높으면 당연히 평균 세대수도 낮다) **만일 전혀 모르는 데이터에 대해 겹치는 변수가 너무 많다면 차원축소가 바람직**

IV. 결과 및 해석

최종 모델과 선택된 변수 설명

Combinatorial Optimization 문제이나, 시간의 제약으로 간단한 Backward Stepwise Regression 방법으로 최종 모델 선택

Variable Selection Strategy: Backward Stepwise Regression

Step 1. Learn Full Model
- RE: include All
- FE: include All

Step 2. Stepwise Variable Elimination, while $|AIC|$ increases
- RE: exclude γ_m s.t. ① $\widehat{sd}(\gamma_m) \approx 0$ or ② $0 \in CI(\widehat{sd}(\gamma_m))$
- FE: exclude β_m s.t. $t.stat(\beta_m) = {\widehat{\beta_m}}/{s.e.(\widehat{\beta_m})}$ small

Step 3. Report final model
- Check diagnostics
- Visualize results

- **Final Model Description**

```
model5 = lmer(FW_t ~ 1 + month_int + TotalHouses+ avg_temp +
                (1 + month_int + age_05 + mem1 + nBusiness|gu),
              data= DF_fit_model5)
faraway::sumary(model5, digits=3)
```

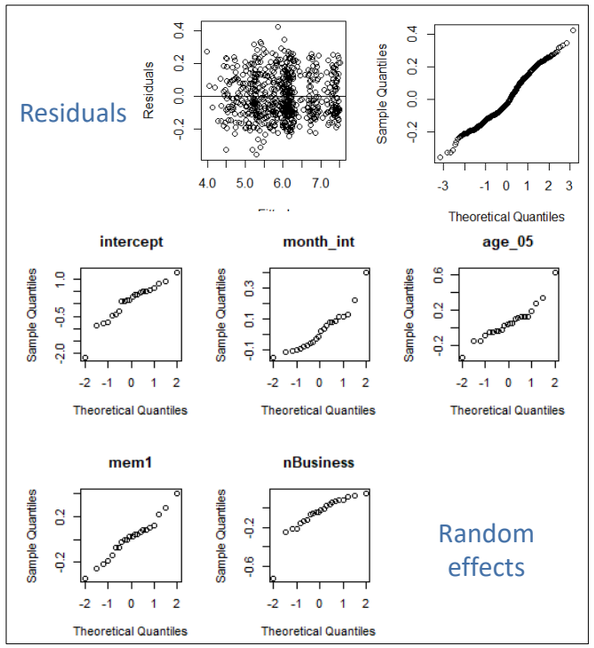
```
Fixed Effects:
            coef.est coef.se
(Intercept) 5.864    0.163
month_int   0.095    0.026
TotalHouses 0.409    0.046
avg_temp    0.034    0.005

Random Effects:
 Groups   Name        Std.Dev
 gu       (Intercept) 0.788
          month_int   0.152
          age_05      0.290
          mem1        0.259
          nBusiness   0.253
 Residual             0.137
---
number of obs: 660, groups: gu, 22
AIC = -484, DIC = -568.1
deviance = -546.0
```

```
t value
 36.027
  3.619
  8.834
  6.284
```

```
> model5_CI_sub %>% round(3)
   2.5 % 97.5 %
1: 0.498  1.056
2: 0.073  0.226
3: 0.038  0.326
4: 0.031  0.396
5: 0.110  0.500
6: 0.130  0.145
```

▲ Parameter Estimates



▲ Diagnostics

$$y_{ij} = \beta_0 + \beta_1 MON_{ij} + \beta_2 TTH_{ij} + \beta_3 TMP_{ij} + \gamma_{0i} + \gamma_{1i} MON_{ij} + \gamma_{2i} AG5_{ij} + \gamma_{3i} MM1_{ij} + \gamma_{4i} NBS_{ij} + \sigma^2$$

Fixed + Random
- **MON:** 시간에 따른 영향 (추세)

Fixed Effects
- **TTH:** 전체 세대수
- **TMP:** 월평균 기온

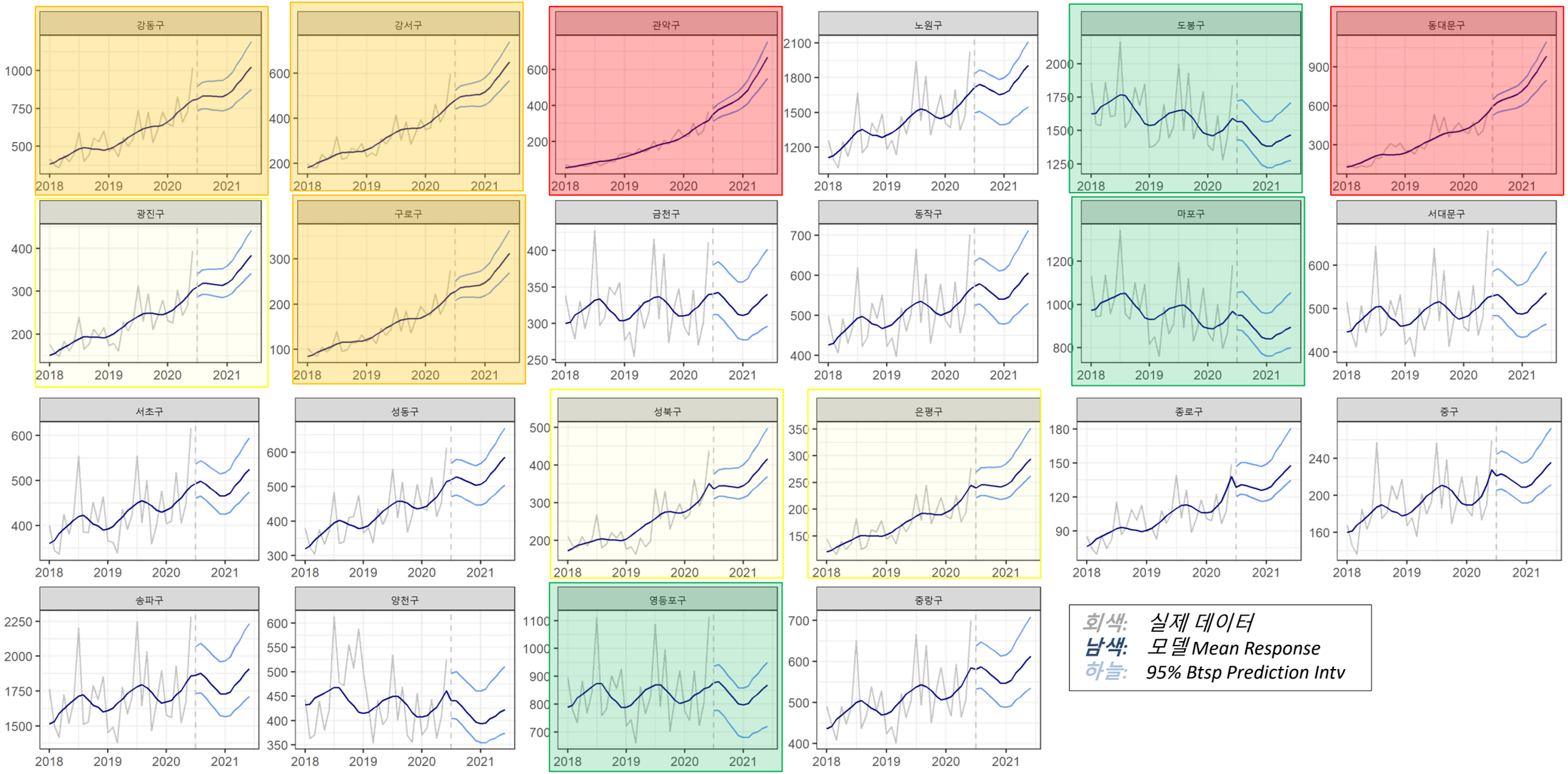
Random Effects
- **AG5:** 5세 이하 인구 비율 (영유아 비율)
- **MM1:** 세대 중 1인 세대 비율
- **NBS:** 점포 수 (지역 경제 활성 정도 연관)

▲ Model Interpretation

IV. 결과 및 해석

예측 기간에 대한 설명변수가 있다면 자치구별로 예측할 수 있다!

모델의 설명변수들은 대부분 18~20년동안 일정한 추세를 보임 (세대수 증가, 1인가구 증가, 영유아 감소, 계절성 기온)
이러한 추세가 지속될 것이라 가정하면, 20년 7월 ~ 21년 6월 월간 배출량을 구별로 예측 가능! **증가 추세가 다른 이유는?**



회색: *실제 데이터*
남색: *모델 Mean Response*
하늘: *95% Btsp Prediction Intv*

Summary

> 50% 증가
- 관악구
- 동대문구

> 20% 증가
- 구로구
- 강서구
- 강동구

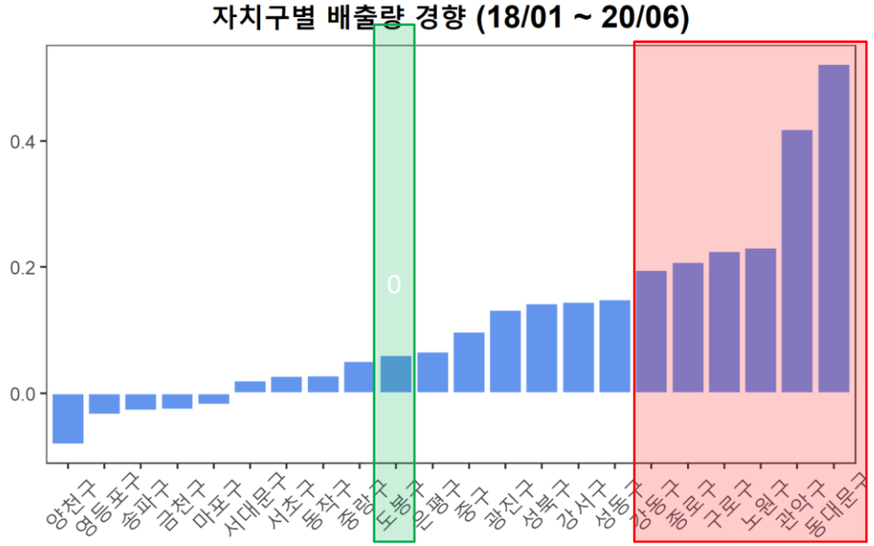
> 15% 증가
- 광진구
- 은평구
- 성북구

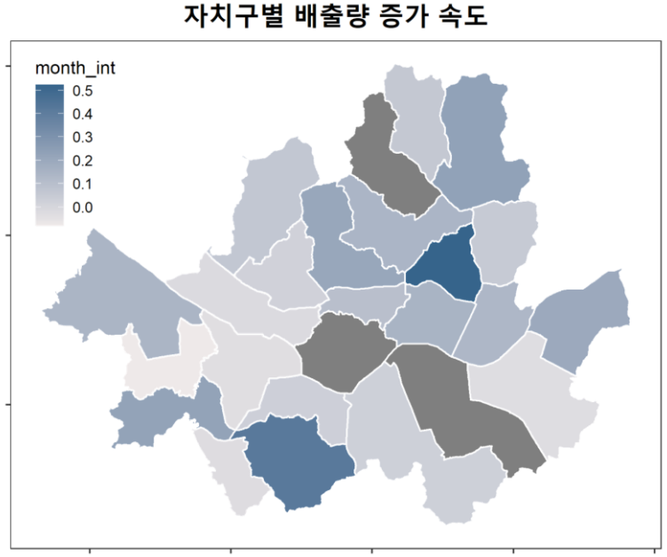
< -10% 감소
- 영등포구
- 마포구
- 도봉구

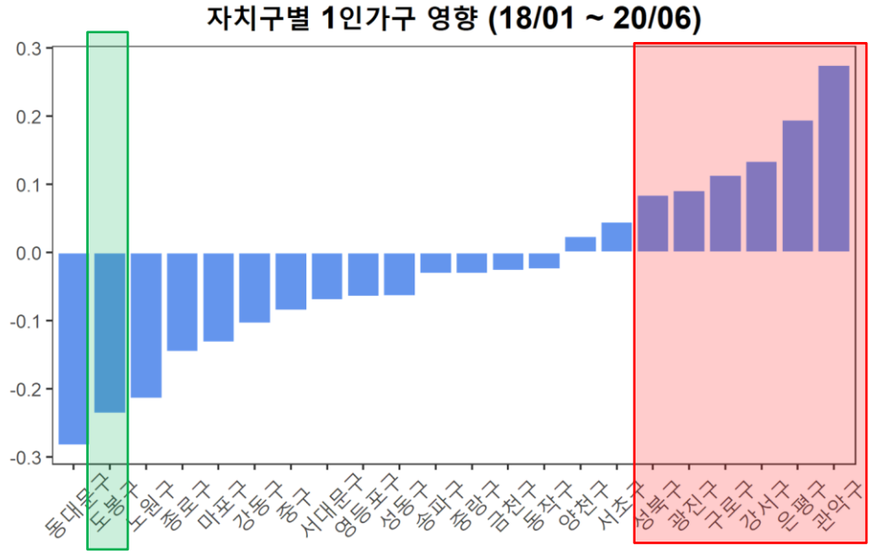
IV. 결과 및 해석

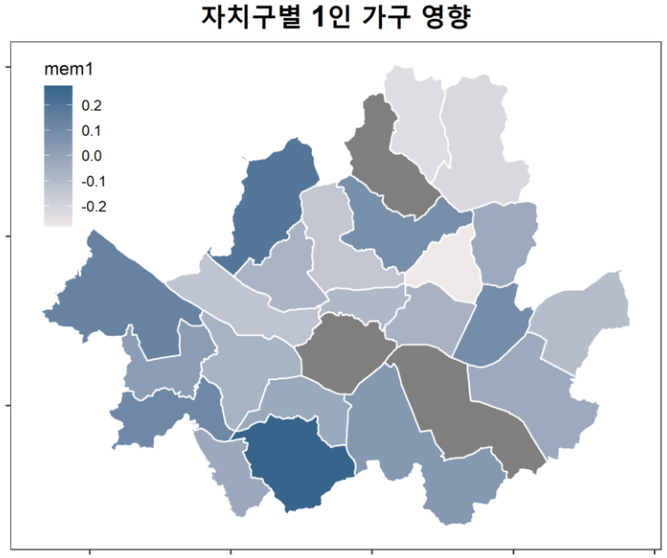
자치구별로 증가 속도가 다른 이유는?

자치구별 배출량 증가 속도의 차이는 Time trend와 1인가구의 Random Effects에서 기인함









Summary

> 50% 증가

- **관악구, 동대문구** 모두 Time trend가 가장 높았음. 모델의 다른 변수가 설명하지 않는 증가 추세로, 자치구별 포착되지 않은 어떤 변수에 기인함
- **관악구**의 경우 1인가구 증가에 따른 영향이 가장 컸음. 1인 가구 증가 추세를 감안할 때 지속적으로 배출량 증가할 것

> 20% 증가 > 15% 증가

- **구로구, 강서구, 광진구, 은평구, 성북구** 모두 1인가구 증가에 따른 영향이 큰 편이었음
- **강동구, 동대문구는** Time trend가 1인 가구 영향을 압도한 것으로 보임

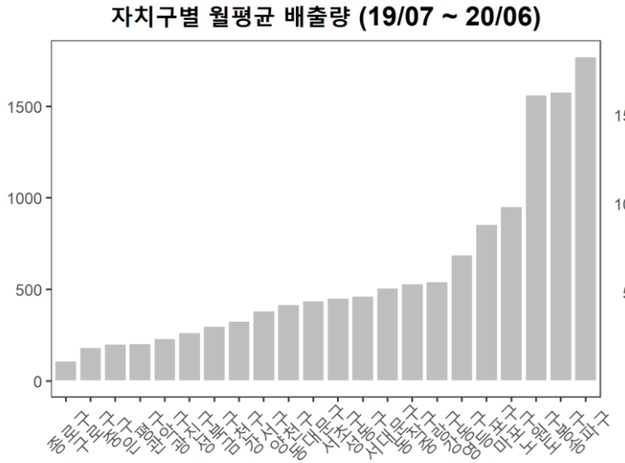
< -10% 감소

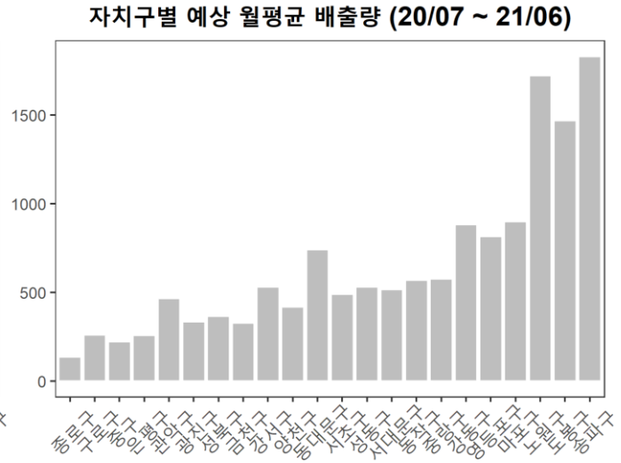
- **영등포구, 마포구는** 모두 Time trend가 감소
- **도봉구**는 Time trend가 증가함에도 불구하고 1인가구 영향에 의해 감소하는 것으로 보임. 도봉구에서는 1인가구 증가 경향이 음식물 쓰레기 배출량 감소로 이어짐

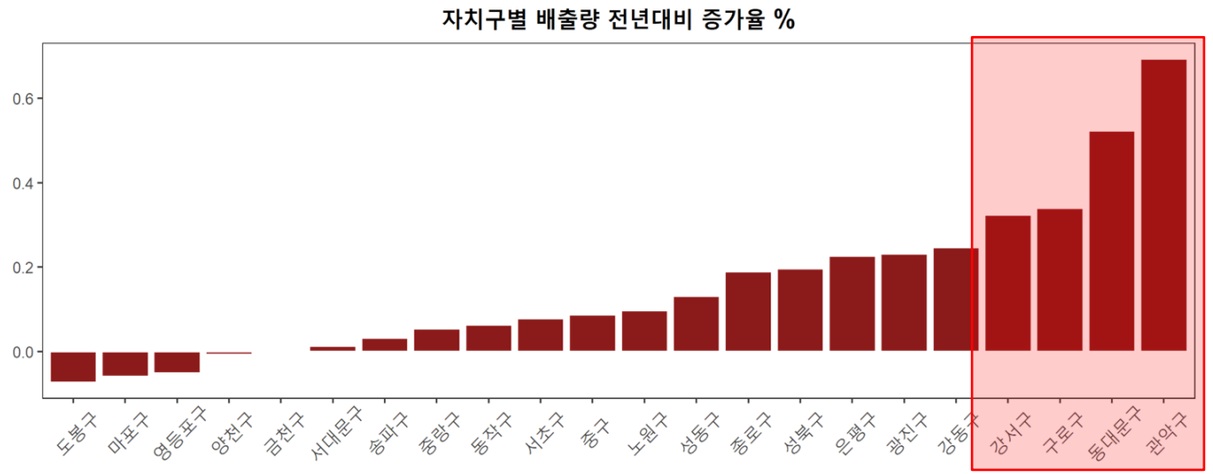
v. 정책 제언

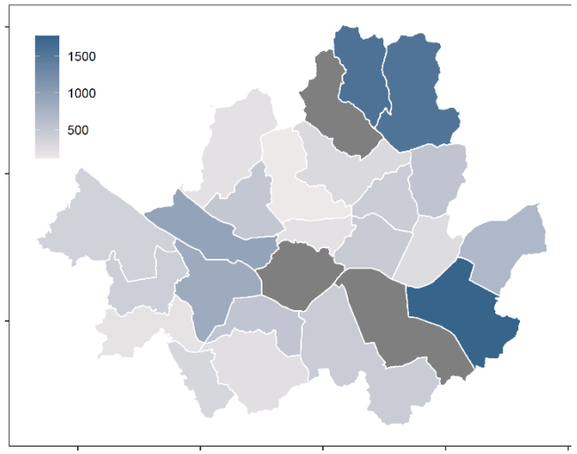
동대문구, 관악구는 음식물 쓰레기 폭탄에 대비하라!

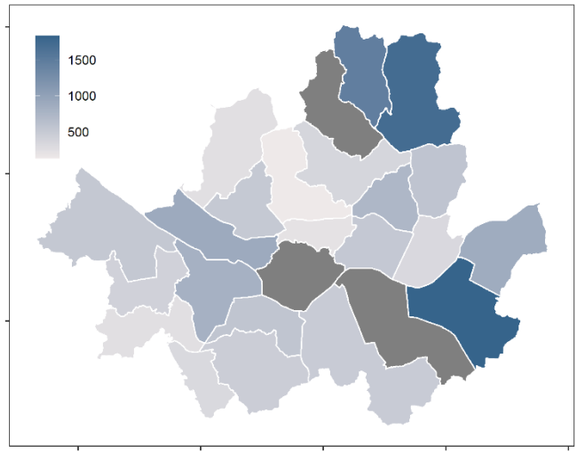
전체적인 수준에 큰 차이는 없으나, 동대문구, 관악구, 구로구, 강서구에 큰 증가폭 예상됨
이들 지역은 상대적으로 배출량 수준이 낮았으나 빠르게 증가하므로, 해당 자치구는 속히 배출량 처리 역량을 확충해야!











한계점

1. **강남구, 강북구, 용산구 제외**
   - 강남, 강북은 기간 내 RFID 종량기 미설치로 인하여 제외 (NA)
   - 용산구는 배출량 수준이 다른 지역에 비해 크게 낮아, 모델 학습 시 Residual의 정규성 가정에 위배됨 (Outlier)
2. **Random Effect 해석의 어려움:**
   - 식당수, 영유아 비율은 육안으로는 배출량 증가 예측과 자치구별 계수를 연관지어 설명하기 어려움
3. **설명변수의 부족:**
   - Random intercept와 Time trend는 데이터가 설명할 수 없는 자치구별 배출량 수준과 증가 추세의 차이를 나타냄. 이를 설명하려면 더 많은 변수가 필요

감사합니다! 질문 받아요!