2023 Generalized Linear Model Course Project

강남구 주 별 아파트 거래량 분석

김상태(2022711082)

김소은(2022811920)

남주연(2022710230)

1. Introduction

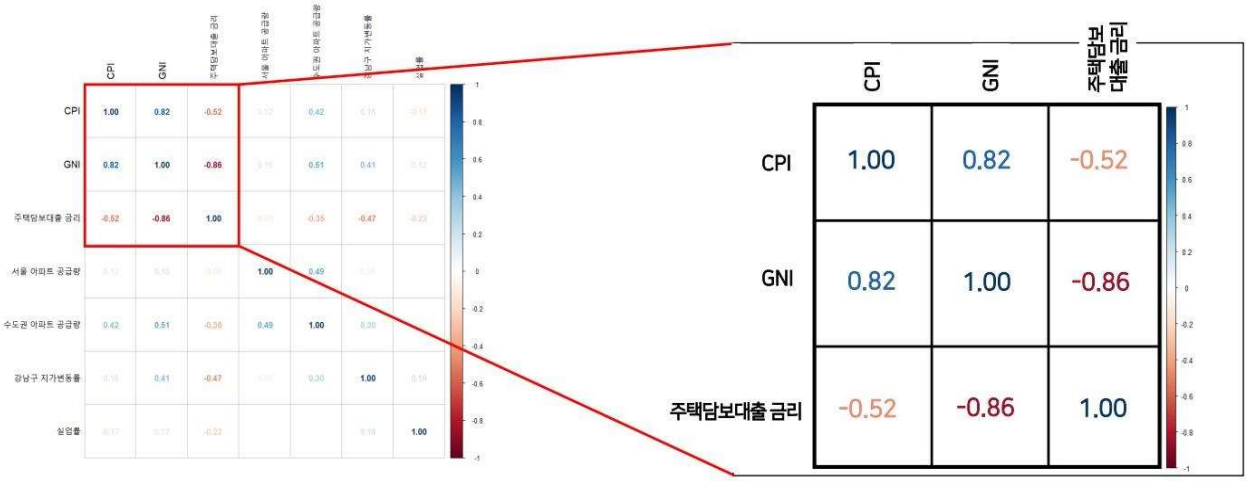
주제를 강남구 주별 아파트 거래량으로 선정하게 된 이유들은 첫번째, 사람들이 살아가면서 가장 중요한 것이 의식주인데 그 중에서 부동산이 유일하게 한정되어 있다. 옷이나 음식은 저렴한 것을 살 수 있지만 부동산은 그렇지 않다. 두번째, 아파트 거래량 데이터를 통해 현재 부동산 시장의 현황을 살펴볼 수 있다. 세번째, 시계열 데이터이다 보니까 date 를 가지고 다른 변수들과 병합이 쉽다. 지역을 강남구로 고른 이유는, 강남구가 종합부동산세 즉 종부세의 대략 26%를 차지한다는 점에서 부동산 시장에서 강남구가 가지는 대표성을 보고 강남구를 고르게 되었다. 분석 방향은 간단한 모형에서 시작을 해서 데이터가 가지는 과대 산포를 설명하는 쪽으로 진행을 할 것이다. 과대 산포를 설명함으로 인해서 추정되는 계수나 모형이 점점 복잡해 질수록 반영이 잘 되는지 살펴볼 예정이다.

간략한 모델링 프로세스는 먼저 강남구 주 별 아파트 거래량과 설명변수의 데이터를 병합을 해서 데이터를 만들 예정이고 전처리와 EDA를 진행한 뒤에 모델링을 GLM with log link, Quasi-likelihood approach, Negative Binomial distribution, Zero Inflated Poisson distribution, Tweedie distribution 으로 5가지 정도 진행을 하고 모형 결과를 비교할 예정이다.

2. Data Description

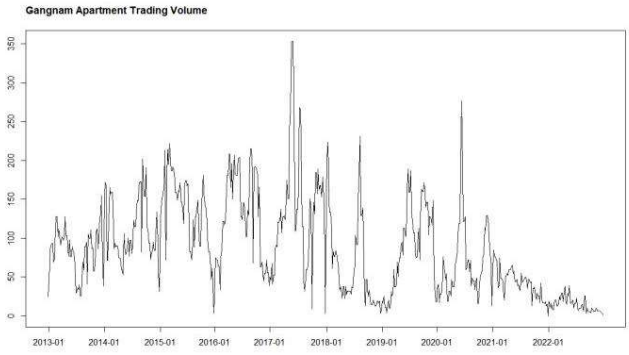
선정한 데이터 기간은 이제 2013년 1주 차부터 2022년 52주 차에서 총 한 10년 정도이다. 624개의 11개의 변수이다. 종속 변수는 강남구 주 별 아파트 거래량이고 독립 변수는 지수변수에 전국 소비자 물가지수(CPI), 1인당 국민총소득(GNI), 실업률이 있고 주택 변수에는 주택담보대출 금리, 강남구 지가 변동률, 서울과 수도권의 아파트 공급량이 있

다. 마지막으로 기타 변수에는 계절 여부와 코로나19 여부가 있다. GLM 은 다중공정성이 있으면 알고리즘 자체가 수렴을 하지 않기 때문에 변수들 간의 다중공정성을 살펴보았다. 아래 그림을 확인하면, CPI와 주택담보대출금리 두 변수와 모두 다중공정성이 높은 GNI 변수 하나를 삭제하였다.

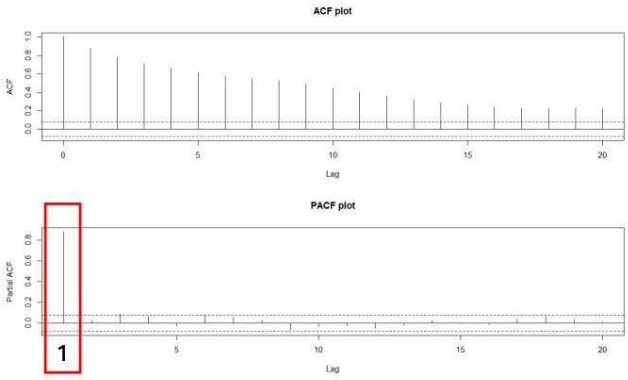


▲ Figure : 독립 변수 별 다중공선성

시차 변수를 파생 변수로 추가하였는데 강남구 주 별 아파트 거래량 이므로, 일종의 시계열 데이터이기 때문에 ACF랑 PACF를 살펴보았다. 아래 그래프를 살펴보면, 느리게 감소하는 ACF 함수 이고 차수1에서 cut-off되는 PACF함수를 보았을 때, 1차수 이전 시차 변수(lagged variable : AR(1))를 파생변수로 고려하였다.



▲ Figure : 강남구 주 별 아파트 거래량 시계열 플롯



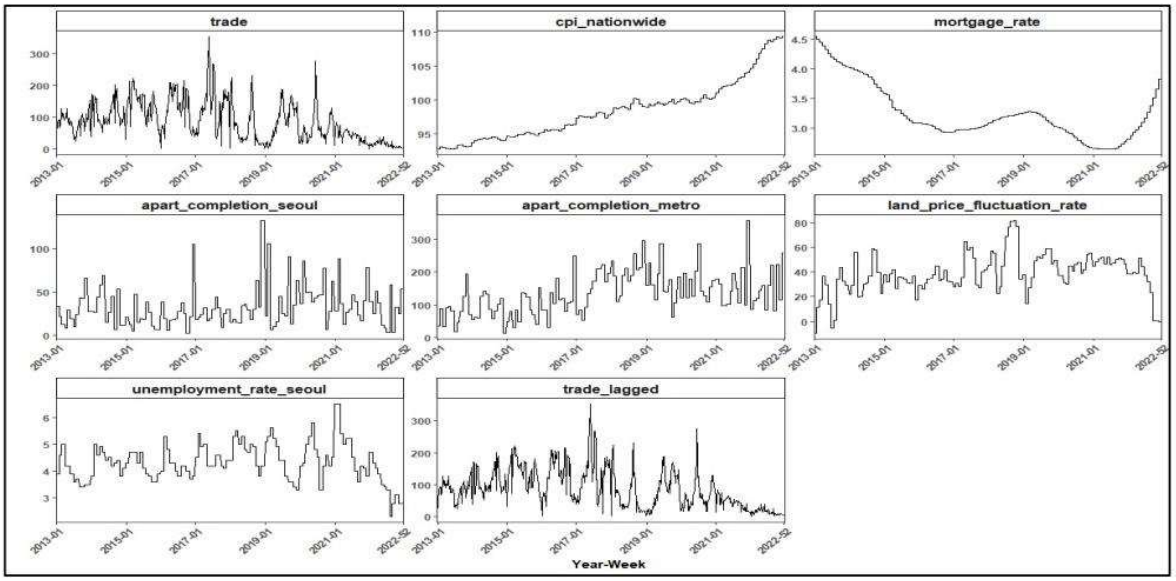
▲ Figure : 강남구 주 별 아파트 거래량 ACF/PACF 플롯

최종 변수를 살펴보면, 맨 마지막에 10번째 변수로 바로 직전 주에 아파트 거래량을 추가하였다. 나머지 변수들의 설명도 살펴볼 수 있다. 범주형 변수에 대해서는, 봄을 베이스 카테고리로 지정하였고, 코로나19 같은 경우는 1이 코로나19가 있는 기간이다.

변수 명	한글 명	종류	설명
Trade	강남구 주 별 아파트 거래량	종속 변수(Count)	0이상의 값만 가지는 Count data
Cpi_nationwide	전국 기준 소비자물가지수	독립 변수(Numeric)	월 단위의 전국 소비자 물가지수
Mortgage_rate	주택담보대출 금리	독립 변수(Numeric)	월 단위의 주택담보대출 금리
Apart_completion_seoul	서울 아파트 공급량	독립 변수(Numeric)	월 단위의 서울 아파트 공급량
Apart_completion_metro	수도권 아파트 공급량	독립 변수(Numeric)	월 단위의 수도권(서울, 경기, 인천) 아파트 공급량
Land_price_fluctuation_rate	강남구 지가변동률	독립 변수(Numeric)	월 단위의 강남구 지가변동률
Unemployment_rate_seoul	서울 실업률	독립 변수(Numeric)	월 단위의 서울 지역 실업률
Season	계절	독립 변수(Character)	봄, 여름, 가을, 겨울 (base category : 봄)
Covid19	코로나19 여부	독립 변수(Character)	1(코로나19 유), 0(코로나 무)
Trade_lagged	시차 변수	독립 변수(Numeric)	바로 직전 주 별 아파트 거래량

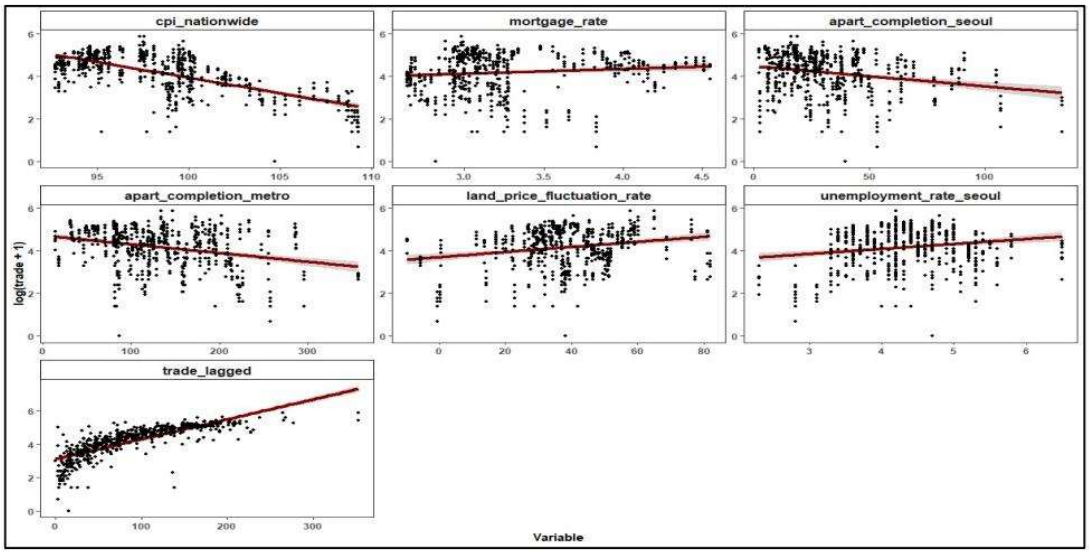
▲ Table : 변수 설명표

EDA 과정을 살펴보겠다. 변수별 추이를 시각화하면, 서울시 아파트 공급량 그래프와 강남구 아파트 거래량(종속변수) 그래프가 서로 어느 정도 비슷한 양상을 띄는 것을 알 수 있다.



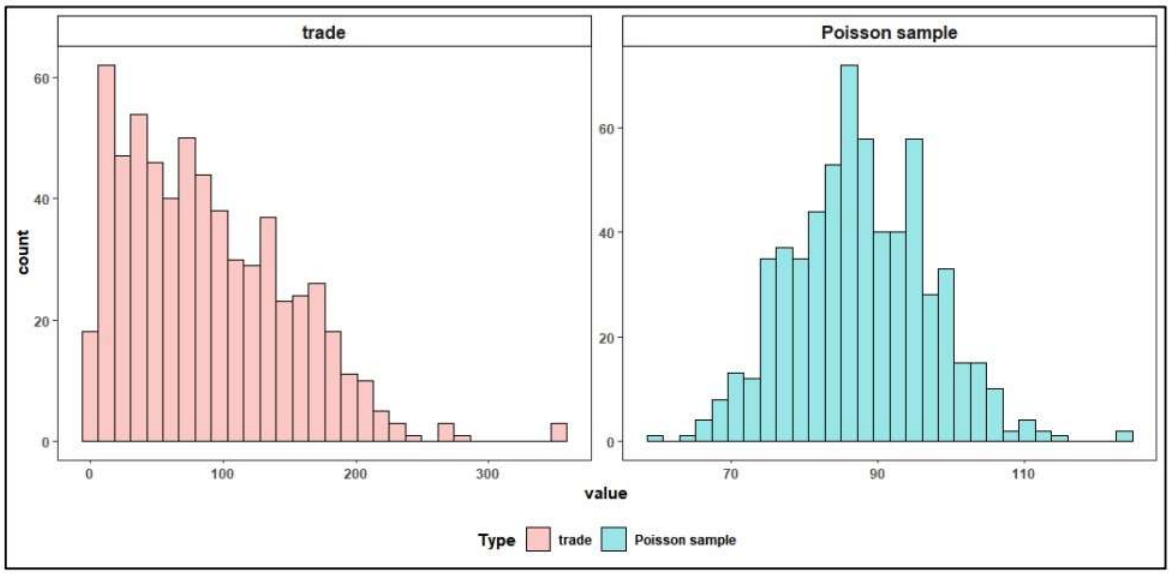
▲ Figure : 변수 별 추이 시각화 플롯

두번째로는 종속변수(아파트 거래량) 와의 관계를 시각화 하여 살펴보았는데, 모든 독립변수들이 log(종속변수)와 선형관계가 있는 것으로 보인다.



▲ Figure : log(종속변수)와 독립 변수 간의 관계 시각화 ** 부정형을 피하기 위해 log(종속변수+1)*

마지막으로 종속변수(아파트 거래량)의 과대 산포(overdispersion)정도를 확인하였다. Sample mean이 87이고 sample variance가 3801로 확인되었다. 오른쪽에 그린 plot이 람다를 87로 했을 때 생성한 sample plot이다. 람다가 크면 클수록 normal approach가 잘 되기 때문에 일종의 조명 shape를 띄어야 하는데 그렇지 못한 과대 산포 특성이 있음을 알 수 있다.



▲ Figure : 아파트 거래량의 Histogram vs Poisson(lambda=87)의 Histogram

3. Modeling

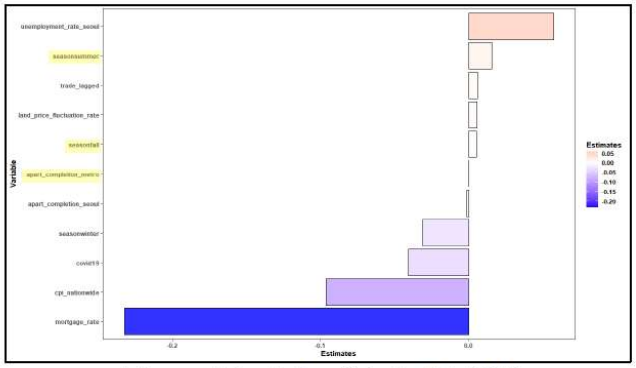
본 프로젝트에서는 아파트 거래량에 다른 종속변수들이 미치는 영향을 알아보기 위해 총 다섯 가지의 모형을 이용한 분석을 진행하였다. Count 데이터를 분석하는데 쓰이는 일반적인 GLM 모형과 데이터 내 과대 산포를 측정할 수 있는 모형들을 사용하였다.

1) GLM with Log Link

Count 데이터를 분석하는 가장 기본적인 모형인 GLM을 사용해 분석을 진행하였다. Random component로 포아송 분포를 가정하고 log link를 사용하였다. 그 결과 실업률, 시차 변수, 지가 변동률이 양의 계수를 가지고, 서울 아파트 공급량, 계절변수 범주 중 겨울, 코로나19 유무, 소비자물가지수, 주택담보대출 금리가 음의 계수를 가지며 유의미한 변수로 나타났다.

```
Call:
glm(formula = trade ~ ., family = poisson(link = "log"), data = select(data_gangnam,
    -c(date_year_week, GNI)))

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                1.355e+01  3.135e-01  43.242  < 2e-16 ***
cpi_nationwide            -9.624e-02  2.932e-03 -32.827  < 2e-16 ***
mortgage_rate             -2.327e-01  1.371e-02 -16.974  < 2e-16 ***
apart_completion_seoul    -1.472e-03  2.597e-04  -5.670 1.43e-08 ***
apart_completion_metro    -8.302e-07  1.015e-04  -0.008   0.9935
land_price_fluctuation_rate 5.707e-03  3.882e-04  14.701  < 2e-16 ***
unemployment_rate_seoul    5.754e-02  9.202e-03   6.253 4.02e-10 ***
seasonsummer               1.566e-02  1.330e-02   1.178   0.2389
seasonfall                 5.307e-03  1.462e-02   0.363   0.7167
seasonwinter              -3.080e-02  1.341e-02  -2.296   0.0217 *
covid19                   -4.049e-02  1.839e-02  -2.202   0.0277 *
trade_lagged               6.080e-03  7.936e-05  76.612  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 27512.4  on 622  degrees of freedom
Residual deviance:  6813.3  on 611  degrees of freedom
AIC: 10558

Number of Fisher Scoring iterations: 4
```

▲ Figure : Summary for GLM



▲ Figure : Estimated coefficients plot of GLM

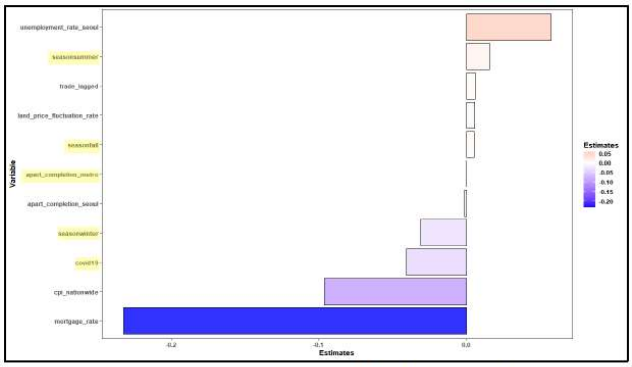
2) Quai-likelihood Approach

Quasi-likelihood 방법은 종속변수에 분포가정 없이 Exponential Dispersion family에 평균이 μ이고 분산을 $\phi\mu$라 가정해 데이터 내 과대 산포를 설명해주는 모형이다. 분석 결과 실업률, 시차 변수와 지가 변동률이 양의 계수를, 서울 아파트 공급량, 소비자 물가지수, 주택담보대출 금리가 음의 계수를 가지며 유의미한 변수로 나타났다. Dispersion

parameter 추정치는 10.8로 데이터의 과대 산포를 설명해주고 있다.

```
Call:
glm(formula = trade ~ ., family = quasipoisson(link = "log"),
    data = select(data_gangnam, -c(date_year_week, GNI)))

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.355e+01  1.028e+00  13.177  < 2e-16 ***
cpi_nationwide          -9.624e-02  9.621e-03 -10.003  < 2e-16 ***
mortgage_rate           -2.327e-01  4.499e-02  -5.172 3.14e-07 ***
apart_completion_seoul  -1.472e-03  8.522e-04  -1.728   0.0845 .
apart_completion_metro  -8.302e-07  3.332e-04  -0.002   0.9980
land_price_fluctuation_rate  5.707e-03  1.274e-03   4.480 8.92e-06 ***
unemployment_rate_seoul  5.754e-02  3.020e-02   1.905   0.0572 .
seasonsummer             1.566e-02  4.364e-02   0.359   0.7198
seasonfall               5.307e-03  4.798e-02   0.111   0.9120
seasonwinter            -3.080e-02  4.402e-02  -0.700   0.4844
covid19                 -4.049e-02  6.034e-02  -0.671   0.5025
trade_lagged             6.080e-03  2.604e-04  23.345  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 10.76942)

    Null deviance: 27512.4  on 622  degrees of freedom
Residual deviance:  6813.3  on 611  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

▲ Figure : Summary for Quasi-GLM



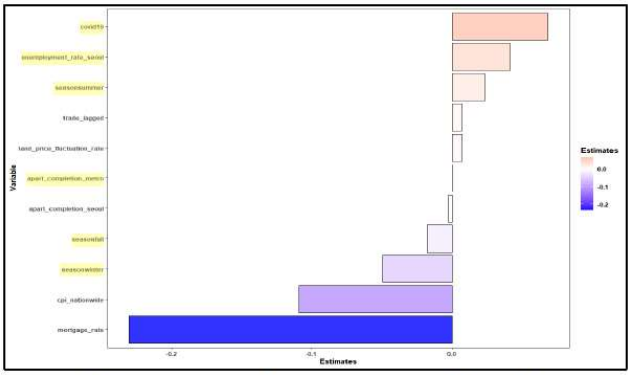
▲ Figure : Estimated coefficients plot of Quasi-GLM

3) Negative Binomial Distribution

일반적으로 분산이 평균보다 큰 Negative Binomial Distribution을 이용한 분석을 진행하였다. 그 결과 시차변수와 지가 변동률이 양의 계수로, 서울 아파트 공급량, 소비자 물가지수, 주택담보대출 금리가 음의 계수를 가지며 유의미한 변수로 나타났다. Dispersion parameter 추정치로는 7.1로 데이터의 과대 산포를 설명해준다.

```
Call:
MASS::glm.nb(formula = trade ~ ., data = select(data_gangnam,
    -c(date_year_week, GNI)), init.theta = 7.063891186, link = log)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              1.464e+01  9.439e-01  15.506  < 2e-16 ***
cpi_nationwide          -1.088e-01  8.336e-03 -13.049  < 2e-16 ***
mortgage_rate           -2.305e-01  4.929e-02  -4.676 2.93e-06 ***
apart_completion_seoul  -2.226e-03  8.750e-04  -2.544    0.011 *
apart_completion_metro  -8.495e-05  3.568e-04  -0.238    0.812
land_price_fluctuation_rate  7.358e-03  1.308e-03   5.625 1.86e-08 ***
unemployment_rate_seoul  4.090e-02  3.095e-02   1.321    0.186
seasonsummer             2.345e-02  5.150e-02   0.455    0.649
seasonfall              -1.766e-02  5.551e-02  -0.318    0.750
seasonwinter            -4.953e-02  4.887e-02  -1.014    0.311
covid19                  6.866e-02  6.060e-02   1.133    0.257
trade_lagged             7.551e-03  3.432e-04  22.005  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(7.0639) family taken to be 1)

    Null deviance: 2598.45  on 622  degrees of freedom
Residual deviance:  662.89  on 611  degrees of freedom
AIC: 5880.4

Number of Fisher Scoring iterations: 1
```

▲ Figure : Summary for NB GLM



▲ Figure : Estimated coefficients plot of NB GLM

4) Zero Inflated Poisson Distribution

데이터의 과대 산포가 종속변수인 아파트 거래량에 0 값이 많아 발생했을 것이라는 가정에서 ZIP 모형을 이용한 분석을 진행하였다. ZIP 모형의 경우, 0 값이 많이 존재하는 count 데이터를 설명하기 위해 포아송 분포와 이항 분포를 혼합해 사용하는 모형이다.

① Random Component : $Y \sim ZIP(\pi, \mu), \quad E(Y) = (1-\pi)\mu, Var(Y) = \mu(1-\pi)(1+\pi\mu)$

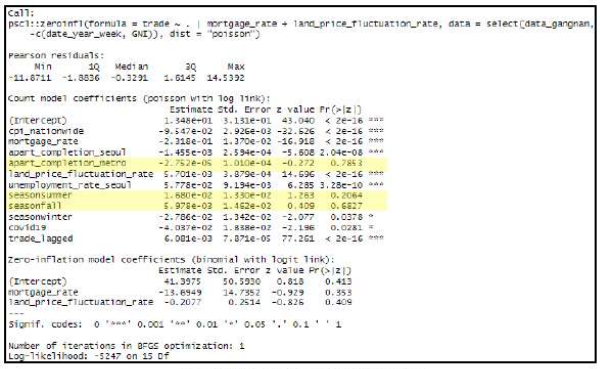
Zero model : $P(Y=0) = \pi + (1-\pi)e^{-\mu}$ *Zero points come from both the binomial and poisson

Non-Zero model : $P(Y=y_i) = (1-\pi)\frac{\mu^{y_i}e^{-\mu}}{y_i!}, \quad y_i = 1,2,3,\dots$

② Systematic Component : $\sum_{j=0}^{p} x_{ij}\beta_j, \quad x_{i0} = (1,\dots,1)'$ *zero model과 non-zero model에 각각 다른 식 적합 가능

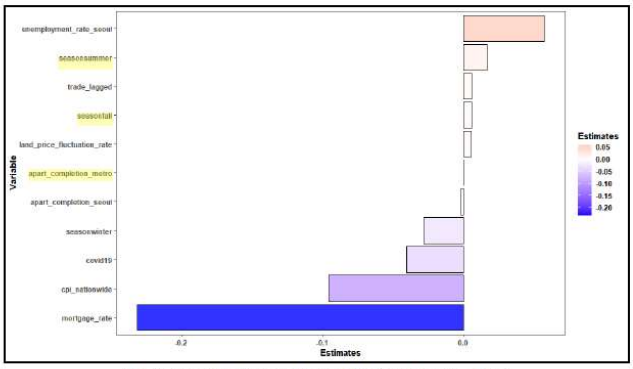
③ Link function : logit link for zero model & log link for non-zero model

이때, 추정된 π 값이 0이 아니어야 포아송 분포를 가정한 GLM 모형과 다른 결과를 보여준다. 분석 결과 실업률, 시차 변수, 지가 변동률이 양의 계수로, 서울 아파트 공급량, 계절변수 범주 중 겨울, 코로나19, 소비자 물가지수, 주택담보대출 금리가 음의 계수를 가지며 유의미한 변수로 나타났다. 하지만 π의 추정치가 0.002로 앞서 사용한 GLM 모형 비슷한 결과를 보여주는 것을 알 수 있다.



```
Call:
pscl::zeroinfl(formula = trade ~ . | mortgage_rate + land_price_fluctuation_rate, data = select(data_gangnam,
    -c(date_year_week, GNI)), dist = "poisson")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-11.8711  -1.8836  -0.3291   1.6145  14.5392

Count model coefficients (poisson with log link):
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               1.348e+01  3.131e-01  43.040  < 2e-16 ***
cpi_nationwide           -9.547e-02  2.926e-03 -32.626  < 2e-16 ***
mortgage_rate            -2.318e-01  1.370e-02 -16.918  < 2e-16 ***
apart_completion_seoul   -1.495e-03  2.194e-04  -5.858 2.04e-08 ***
apart_completion_metro   -2.752e-05  1.010e-04  -0.272   0.7853
land_price_fluctuation_rate 5.701e-03  3.879e-04  14.696  < 2e-16 ***
unemployment_rate_seoul   5.778e-02  9.194e-03   6.285 3.28e-10 ***
seasonsummer              1.680e-02  1.192e-02   1.283   0.2064
seasonfall                5.978e-03  1.162e-02   0.515   0.6827
seasonwinter             -2.786e-02  1.342e-02  -2.077   0.0378 *
covid19                  -4.037e-02  1.888e-02  -2.196   0.0281 *
trade_lagged              6.001e-03  7.871e-05  77.261  < 2e-16 ***

Zero-inflation model coefficients (binomial with logit link):
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                41.3975    50.5930   0.818    0.413
mortgage_rate             -15.6949    14.7352  -0.979    0.553
land_price_fluctuation_rate -0.2077     0.2514  -0.826    0.409
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 1
Log-likelihood: -5247 on 15 Df
```

▲ Figure : Summary for ZIP



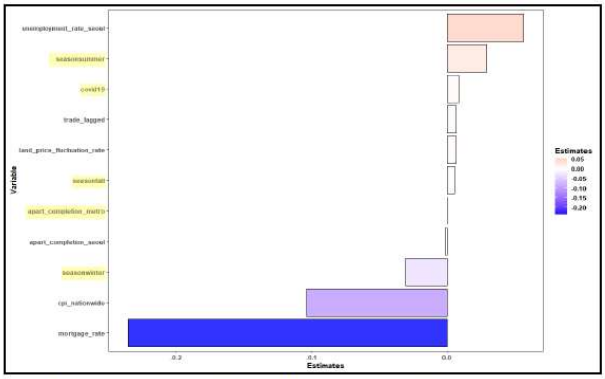
▲ Figure : Estimated coefficients plot of ZIP

5) Tweedie Distribution

Tweedie Distribution 모형은 Quasi-likelihood 모형의 일반화된 형태라고 볼 수 있다. Quasi-likelihood Approach와 동일하게 종속변수에 대한 분포 가정없이 Exponential Dispersion Family에 평균은 μ, 분산은 $\phi\mu^p$라 가정한 모형이다. 여기서 ϕ는 dispersion parameter이고, p는 Tweedie power parameter을 의미한다. p의 추정값이 1인 경우 Quasi-likelihood 모형과 동일한 결과를 보여준다. 분석 결과 실업률, 시차 변수와 지가 변동률이 양의 계수를, 서울 아파트 공급량, 소비자 물가지수, 주택담보대출 금리가 음의 계수를 가지며 유의미한 변수로 나타났다. ϕ와 p는 각각 1.4와 1.6로 추정되었다.

```
Call:
glm(formula = trade ~ ., family = tweedie(var.power = fit$p.max,
    link.power = 0), data = select(data_gangnam, -c(GNI, date_year_week)))

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.422e+01  9.619e-01  14.779  < 2e-16 ***
cpi_nationwide           -1.040e-01  8.745e-03 -11.893  < 2e-16 ***
mortgage_rate            -2.351e-01  4.583e-02  -5.130 3.89e-07 ***
apart_completion_seoul   -1.759e-03  8.585e-04  -2.049   0.0409 *
apart_completion_metro   -5.461e-05  3.422e-04  -0.160   0.8733
land_price_fluctuation_rate  6.622e-03  1.274e-03   5.197 2.77e-07 ***
unemployment_rate_seoul   5.597e-02  3.021e-02   1.853   0.0644 .
seasonsummer              2.902e-02  4.698e-02   0.618   0.5369
seasonfall                5.524e-03  5.127e-02   0.108   0.9142
seasonwinter             -3.086e-02  4.595e-02  -0.672   0.5020
covid19                   8.846e-03  5.973e-02   0.148   0.8823
trade_lagged              6.691e-03  2.908e-04  23.006  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Tweedie family taken to be 1.583451)

    Null deviance: 4161.2  on 622  degrees of freedom
Residual deviance: 1040.5  on 611  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

▲ Figure : Summary for Tweedie GLM



▲ Figure : Estimated coefficients plot of Tweedie GLM

4. Conclusion

아파트 거래량에 영향을 미치는 종속변수들에 대해 알아보기 위해 GLM, Quasi-likelihood, Negative Binomial Distribution, Zero Inflated Poisson Distribution 그리고 Tweedie Distribution의 총 다섯가지 모형을 사용한 분석을 진행하였다. 분석 결과는 다음과 같이 정리할 수 있다.

1) Comparison – Coefficients of Models

다섯가지 모형 모두 대체적으로 비슷한 방향의 추정계수를 주었지만, Negative Binomial 모형의 경우 계절변수 범주 중 가을과 코로나 19 변수에서 다른 모형들과는 반대 부호의 추정계수를 주었다. 수도권의 아파트 공급량, 계절변수 범주 중 여름과 가을은 모든 모형에서 유의미하지 않다고 나왔다. 소비자물가지수, 주택담보대출 금리, 서울의 아파트 공급량, 강남 지가 변동률과 거래 시차 변수는 모든 모형에서 유의미한 변수로 나왔다.

시차 변수와 강남의 지가 변동률의 추정치는 양의 값을 보이고 있다. 이는 이전 주차의 아파트 거래량이 많으면 다음 주차의 거래량 또한 많음을 의미하고, 지가 변동률이 상승하게 되면 소비 심리에 의해 거래량이 증가함을 의미한다. 소비자물가지수, 주택담보대출 금리와 서울의 아파트 공급량은 모든 모형에서 음의 계수를 갖는다. 이는 소비자물가지수와 주택담보대출의 금리가 상승하거나 서울의 아파트 공급량이 증가하게 되면, 경기 불황 및 아파트 값이 떨어질 것이라는 소비 심리에 의해 아파트의 거래량이 줄어듦을 의미한다.

2) Comparison – Model Performance

모형들의 성능을 비교하면, Residual Deviance 기준으로 NB, Tweedie 모형이 좋은 성능을 보여줬다. AIC 기준으로는 과대 산포를 설명할 수 있는 NB, Tweedie, Quasi-likelihood 모형 순으로 좋은 성능을 보여줬다. Tweedie와 Quasi-likelihood의 모형의 경우, 분포가정이 없어 AIC 값을 구할 수 없는데, 모형의 성능 비교를 위해 $AIC^* = Deviance + 2 \times (no.of\ parameters)$을 사용해 구한 값을 사용하였다.

ZIP의 경우, 예상했던 바와 달리 성능이 좋지 않았다. 이는 ZIP 모형을 적용하기에 아파트 거래량이 0인 경우, 즉 아파트 거래가 특정 주차에 한 번도 발생하지 않은 경우가 너무 적게 존재했기 때문이라 판단했다. 실제로 거래량을 좀 더 상세히 살펴본 결과 거래량이 0인 경우가 한 번 밖에 없었다.

대체적으로 데이터의 과대 산포를 설명해 주는 모형이 더 좋은 성능을 보여주고 있다. GLM과 ZIP 모형의 경우, 데이터의 과대 산포를 설명해주지 못한다. 나머지 세 모형의 경우 과대 산포를 설명해 주는데 데이터가 가지는 과대 산포 정도가 너무 커 모든 모형이 이를 정확하게 잡아내지는 못하였다. 모든 모형이 Sample Variance 값보다 작은 값으로 추정하였으며, NB, Tweedie 그리고 Quasi 모형 순으로 크게 추정하였다.

3) Future Study

앞선 분석 결과들을 바탕으로 추후 다양한 연구가 진행 가능할 것 같다. 우선 데이터의 과대 산포가 큰 것이 데이터 내 이상치가 너무 많기 때문이라 생각해 이상치들을 제거 후 다시 모형 적합을 진행해 볼 수 있다. 또한 프로젝트에서 사용한 모형 외에도 과대 산포를 설명할 수 있는 다른 모형들도 고려해 볼 수 있을 것이다. 아파트 거래량이 시계열 데이터인 것을 고려한다면 다른 시계열 분석 기법을 활용한 분석도 가능하다. 또한 강남구 외에도 서초구, 송파구 등 인접 지역의 거래량이나 공간 정보를 설명 변수로 추가해 다차원 회귀 모형도 고려할 수 있다. 추후 이와 같은 다양한 추가 분석이 이루어진다면 아파트 거래량과 관련해 좀 더 의미 있는 분석이 진행될 것 같다.