기상데이터를 활용한
서울시 공공자전거 따릉이 수요 예측

2013121145 이정현

2015116034 이송아

2015122039 김하늘

2015122057 최종문

<목 차>

1. 서론

2015년 이전의 서울시의 모습과 현재의 모습, 작년과 현재의 연세대의 모습을 찬찬히 떠올려보면 눈에 띄는 차이를 느낄 수 있다. 어느 순간 인도에는 자전거 거치대와 함께 일명 '따릉이' 라고 불리는 자전거가 놓이게 되었고, 연세대 각 건물들 앞에는 kickgoing이라고 쓰여 있는 전동킥보드와 전기 자전거가 다수 놓여있는 것을 볼 수 있다. 이는 환경오염의 대안으로 새롭게 떠오르고 있는 이른바 '공유경제'의 일환으로, 공유경제란 한번 생산된 제품을 여럿이 공유해 쓰는 협력소비를 기본으로, 물품은 물론 생산 설비나 서비스 등을 개인이 소유할 필요 없이 필요한 만큼 빌려 쓰고, 자신이 필요 없는 경우 다른 사람에게 빌려주는 행위를 의미한다. 최근에는 이것이 경기침체와 환경오염에 대한 대안을 모색하는 사회운동으로 확대되어 사용되고 있으며, 서울시에서도 이에 부합하여 서울시의 교통체증과 대기오염, 고유가 문제를 해결하고 건강한 사회 및 시민들의 삶의 질을 높이고자 공공자전거 서비스 사업, 즉 따릉이 사업을 실시하였다.

서울시의 공공자전거 사업은 2010년 11월 시범운영을 시작하여 2015년 9월까지 확대 및 시범운영을 계속해오다 2016년 3월 5대의 거점 및 2000대의 자전거와 함께 공단 운영을 개시하며 본격화되었다. 지난 3년간 서울시의 따릉이 사업 규모는 점차 확대되었으며 2019년 4월 30일 기준으로는 1540여개의 대여소와 20000개의 자전거, 19293개의 거치대가 운영되고 있다

현재 서울시 뿐만 아니라 광주시에서도 공공자전거 대여 서비스 '따랑께'를 준비하고 있으며, 지방자지단체들 뿐만 아니라 많은 스타트업 기업들이 스마트 모빌리티 대여 서비스를 속속 출시하고 있다. 이런 사업을 준비하는 정부기관 혹은 기업들에게 해당 서비스의 정확한 수요 예측은 큰 도움이 될 것이다. 따라서 우리는 현재 이용자 수가 가장 많고 데이터가 잘 축적되어 있는 서울시 공공자전거 따릉이의 데이터를 분석하여 따릉이의 수요를 예측하는 연구를 진행하기로 했다.

따릉이의 이용횟수에 영향을 미치는 요인으로 가장 먼저 떠오른 것은 단연 '날씨' 변수이다. 비가 오면 자전거를 타지 않고, 기온이 너무 낮거나 높아도 자전거를 타지 않을 가능성이 높다는 것은 유추 가능한 사실이다. 가림막 없이 그냥 이용해야 하는 자전거의 특성상, '타다'와 같은 자동차 공유 서비스와 달리 기상 요인이 영향을 미치는 바가 클 것이라고 생각하여 날씨변수를 고려하기로 하였다. 기상 데이터는 일별자료를 활용했으며 따릉이 사업이 충분히 성장했다고 볼 수 있는 2018년의 자료를 활용하기로 하였다.

따릉이 이용량에 영향을 미치는 요인으로는 기상요인 이외에도 이용자의 나이, 일별 신규 회원 가입수, 사람들의 환경 및 여가에 대한 관심도, 따릉이의 대여소 수 등이 있다. 하지만 우리는 날씨와 휴일 변수만을 고려하여 분석을 진행하기로 하였는데 그 이유는 다음과 같다.

먼저 날씨와 휴일 변수만을 이용하여 회귀분석을 진행하였을 때에도 회귀식의 설명력을 의미하는 R^2가 충분히 높게 나왔다는 사실이다. 이외의 변수들을 이용하는 데에는 한계가 존재했다. 첫째로 나이 변수의 경우, 2018년 7월 이후부터는 따릉이를 이용하는 회원 및 일반 회원의 성별 및 생년월일을 기재하는 것이 선택사항으로 바뀜에 따라 데이터 자료에 결측값이 많아졌다. 또한 우리가 분석하는 것은 일별 따릉이 이용횟수라는 시계열 자료인데, 각 연령대의 수가 일별로 크게 변동하는 것이 아니기 때문에 나이 요소를 고려하지 않아도 된다고 생각하였다. 둘째로 사람들의 여가 및 환경에 대한 관심도는 관찰 불가능한 변수이기 때문에 분석에서 제외하기로 하였다. 물론 이 관심도를 간접적으로 측정할 수 있는 요소로 따릉이의 회원으로 등록하는 일별 신규 회원수나 따릉이 정기권 일별 구매자 수 등을 이용할 수는 있겠으나 이 또한 자료가 존재하지 않고 따릉이는 회원과 비회원, 정기권고 일회용권을 가진 사람 모두 이용가능하기 때문에 분석에서 제하기로 하였다. 셋째로 따릉이 대여소 수와 같은 경우에는, 일별 자료가 아닌 월별자료만 나와 있고, 월별 수를 봤을 때 월별로 대여소 수가 크게 달라지지 않기 때문에 일별 차이는 더욱 미세할 것이라고 생각하여 고려하지 않기로 하였다. 실제로 2019년 1월부터 4월까지 대여소는 1537개에서 1540개로 단 세 곳만 증가하였음을 확인할 수 있다. 따릉이의 대여소 설치는 주로 연초에 정해진 확대방안 계획에 따라 시행되고, 중간에 대여소를 따로 설치하는 경우는 시민의 설치 의견을 수렴한 후 서울시에서 마련한 기준에 부합하는지를 확인하는 조사과정을 거쳐 토지 및 전기 사용 협의 후 최종 결정되는 긴 과정을 거치기 때문에 일별 이용량에 미치는 영향은 더욱 작을 것이라고 생각하였다. 마지막으로 주위의 회사 개수 등은 지역별 분석 및 비교가 아닌 일별 총 이용량을 분석하는 우리의 시계열 자료에는 적합하지 않다고 생각하여 제외하였다.

따라서, 따릉이의 일별 총 이용량에 대한 설명변수로 날씨 변수를 선택하였다. 이에 날씨 변수 이외에 기준이 명확하고 쉽게 파악할 수 있는 휴일이라는 변수를 추가하여 분석하기로 하였다. 따릉이를 타는 목적이 출근용이냐 휴일에 휴식 및 운동을 위한 용도냐에 따라 그 이용량이 달라지겠지만 '휴일'이라는 변수가 따릉이 이용에 충분히 중요한 역할을 할 것이라고 생각 하였기 때문이다. 다시 한 번 필요한 변수를 뺀 위험이 있을 수도 있다는 것을 충분히 인지하고 진행하였으며, 날씨와 휴일 변수만을 이용하고도 R^2가 높게 나와 분석을 그대로 진행하였음을 밝혀 두고 싶다. 앞으로의 분석은 모두 자료의 부족 및 기타 이유로 제외하게 된 기타 변수들이 꼭 필요한 변수들이 아님에 따라 편향(bias)이 생기지 않았다는 전제하에 진행된 것임을 분명히 해 두겠다.

2. 본론

1. 모형설정

(1) 변수설정

데이터에는 종속변수인 count를 제외하고 총 21개의 변수가 있다. 21개의 변수 중에는 평균기온, 최고기온처럼 다중공선성이 강하게 나타날 것이라고 생각되는 변수들이 여럿 있어 주어진 변수를 다 사용하는 것은 비효율적인 방법이라 판단하였고, 따라서 21개의 변수 중 종속변수를 설명하는 데 가장 유의미할 것으로 예상되는 변수만을 설명변수로 지정하기로 하였다.

선택 과정은 다음과 같다. 21개의 변수를 각각 설명변수로 사용하여 21번의 단일 회귀 모형을 실행했고 각 계수의 t값을 비교해보았다. 같은 범주(ex.온도, 강수량)에 속하는 변수들이 여러 개 존재할 경우 그 중 t값이 가장 높은 변수를 최종 변수로 선택하였고 p-value가 0.05보다 클 경우 종속변수에 유의미한 영향을 미치지 못할 것이라 판단하여 제외하였다. 단, 단일 회귀 모형에서의 결과는 다중 회귀 모형에서 계수 간의 영향력이 중복되어 다른 결과가 나올 수 있음을 인지하고 분석을 진행할 필요가 있다.

최종적으로 선택된 변수들은 다음과 같다.

변수명	변수 설명
count	일별 서울시 공공자전거 따릉이 총이용횟수
mtemp	최고기온(섭씨)
rtime	계속강수시간(시간)
awind	평균풍속(m/s)
ahumid	평균상대습도(%)
tsunamhour	1시간 최다일사량(MJ/m2)
holiday	휴일여부 (0:평일, 1:주말 및 공휴일)

(2) 분석결과

R을 이용한 OLS 다중회귀분석을 한 결과, 다음과 같은 결과를 얻었다.

```
Call:
lm(formula = count ~ mtemp + rtime + awind + ahumid + tsunamhour +
    factor(holiday))

Residuals:
   Min     1Q Median     3Q    Max
-20524  -7773  -1089   6758  31153

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      11240.10    3854.53   2.916  0.00377 **
mtemp              817.77      79.33  10.309  < 2e-16 ***
rtime            -1173.42     176.99  -6.630 1.24e-10 ***
awind            -3033.85     975.82  -3.109  0.00203 **
ahumid              85.71      60.19   1.424  0.15529
tsunamhour        2714.64    1059.65   2.562  0.01082 *
factor(holiday)1 -2669.90    1131.60  -2.359  0.01884 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9954 on 358 degrees of freedom
Multiple R-squared:  0.6414, Adjusted R-squared:  0.6354
F-statistic: 106.7 on 6 and 358 DF,  p-value: < 2.2e-16
```

먼저 t값을 살펴보면 절편항, 종속변수항 6개, 총 7개 중에서 ahumid를 제외한 6개 항이 2 이상의 절대값 t값을 가지고 있는 것을 확인 할 수 있다. t값의 절대값이 2이상이면 유의한 것으로 판단하기로 하자. 그리고 그 절대값이 크면 클수록 독립변수와 종속변수 사이의 인과관계가 더 유의하다. 가장 유의한 것은 mtemp, rtime으로 각각 10.309, 6.630의 t 값을 갖는다. 그 다음으로는 awind, 절편항이 3.109, 2.916으로, 어느정도 유의하다. 마지막으로 tsunamhour 과 holiday는 t값이 2.562, 2.359으로 조금 덜 유의하다. ahumid의 경우 개별회귀분석에서는 유의미하다는 결과가 있었으나 다중회귀분석에서는 그 정도가 상당히 약화되었음을 알 수 있다. R-square의 경우 0.6414로 상당히 높은 수치를 보인다. 이는 따릉이 이용횟수의 64% 가량이 위에서 설정했던 6개의 변수로 설명될 수 있음을 시사한다.

추정치를 이용하여 추정된 회귀식을 세우면 다음과 같다.

$$\hat{Y}_{count} = 11240.1 + 817.77X_{mtemp} - 1173.42X_{rtime} - 3033.85X_{awind} + 85.71X_{ahumid} + 2714.64X_{tsunamhour} - 2669.9X_{holiday}$$

각각 추정된 회귀 계수는 따릉이 이용횟수에 원인변수가 어떻게 영향을 주는지를 보여준다. 섭씨 기온이 1도 증가 했을 때 따릉이 이용횟수는 817.77회만큼 증가한다. 계속강수시간이 1시간 늘어나면 따릉이 이용횟수는 1173.42회 만큼 감소한다. 평균풍속이 1m/s 증가하면 이용횟수는 3033.85회 만큼 감소한다. 평균상대습도가 1%증가하면 따릉이 이용횟수는 85.71회 증가한다. 1일

최다일사량이 1시간 늘어나면 따릉이 이용횟수는 2714.64 회만큼 증가한다. 휴일인 경우에는 따릉이 이용횟수가 2669.9회 만큼 감소한다

2. 표준가정의 해제

(1) 이분산성

(1)-1 이분산성의 진단

선형 회귀 모형에서 우리는 종속변수 값의 분산이 설명변수의 값과 상관없이 고정된 값을 가져야 한다고 가정한다. 하지만 실제 데이터에서는 독립변수의 변화에 따라 종속변수의 분산도 커지는 이분산성(hetroskedasticity) 문제가 발생한다. 데이터에 이분산성이 있는 경우 최소제곱법을 이용해 구해진 추정량은 여전히 불편추정량(unbiased estimator)이지만 더 이상 효율적이지 않다.

```
Call:
lm(formula = count ~ mtemp + rtime + awind + ahumid + tsunamhour +
    factor(holiday))

Residuals:
   Min     1Q Median     3Q    Max
-20524  -7773  -1089   6758  31153

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      11240.10    3854.53   2.916  0.00377 **
mtemp              817.77      79.33  10.309  < 2e-16 ***
rtime            -1173.42     176.99  -6.630 1.24e-10 ***
awind            -3033.85     975.82  -3.109  0.00203 **
ahumid              85.71      60.19   1.424  0.15529
tsunamhour        2714.64    1059.65   2.562  0.01082 *
factor(holiday)1 -2669.90    1131.60  -2.359  0.01884 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9954 on 358 degrees of freedom
Multiple R-squared:  0.6414, Adjusted R-squared:  0.6354
F-statistic: 106.7 on 6 and 358 DF,  p-value: < 2.2e-16
```

기본 모형의 회귀분석 결과는 다음과 같다. 기본 모형에 이분산성이 있는지를 검정하기 위해 White test를 실시했다. 통계프로그램 R에서는 Breusch-Pagan test 함수에 변수를 지정해주지 않으

면 White test가 자동으로 실행된다. White test는 설명변수의 개수가 k개, 데이터의 개수가 n개일 때, k(k+1)/2>n 이면 실행 될 수 없지만 우리는 2018년 전체(n=365) 자료를 이용하여 데이터 개수가 충분하였기 때문에 이를 진행하였다.

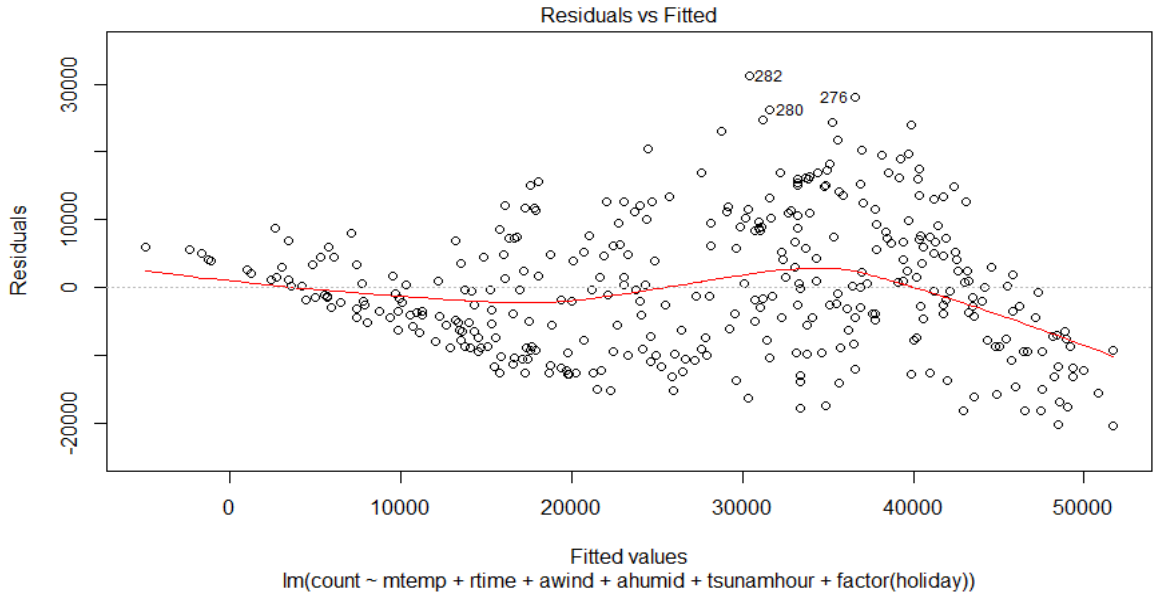
```
 studentized Breusch-Pagan test

data:  ll
BP = 37.584, df = 6, p-value = 1.355e-06
```

분석 결과, p값은 0.00000135으로 이분산성이 발생하였음을 확인할 수 있다.

(1)-2 이분산성의 해결

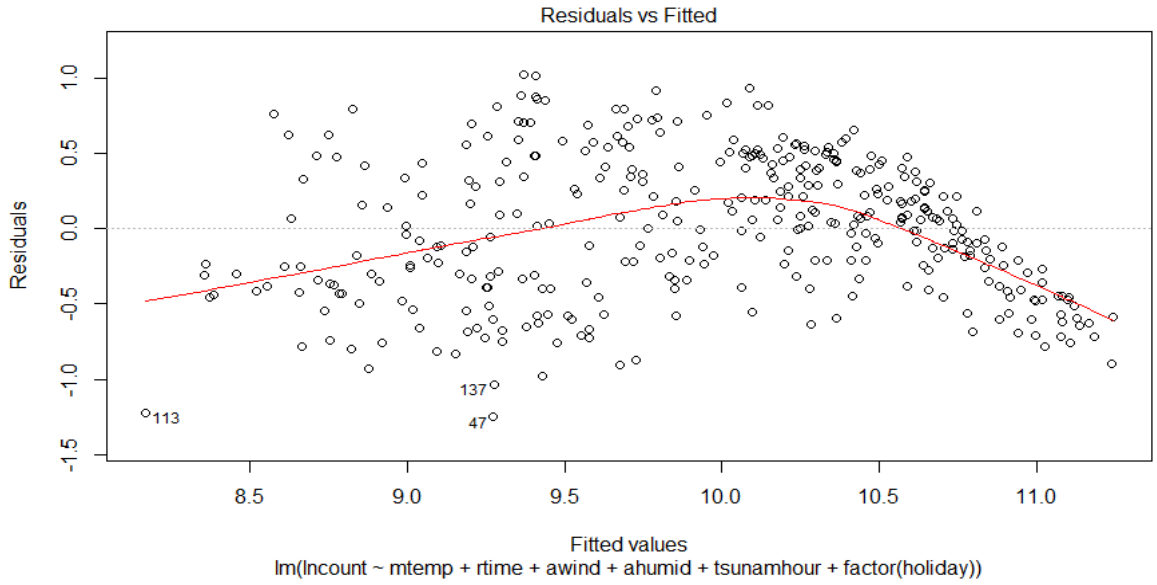
(1)-2-1 종속변수의 로그변환(transformation)

이분산성 문제는 종속변수를 로그변환한 트랜스로그 모형을 사용하면 해결되는 경우가 있다. 이분산성의 문제를 확인하기 위해 기본모형의 residual plot을 그려보았다.



다음과 같이 분산이 고르지 않게 분포함을 알 수 있다.

문제를 해결하기 위해 종속변수를 로그변환하여 새로운 회귀모형을 도출했다. 로그변환한 회귀모형의 residual plot과 회귀분석 결과는 다음과 같다.




```
Call:
lm(formula = lncount ~ mtemp + rtime + awind + ahumid + tsunamhour +
    factor(holiday))

Residuals:
    Min      1Q  Median      3Q     Max
-1.2518 -0.3625  0.0196  0.3903  1.0212

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       9.040235   0.181391  49.838  < 2e-16 ***
mtemp             0.043650   0.003733  11.693  < 2e-16 ***
rtime            -0.069313   0.008329  -8.322 1.85e-15 ***
awind            -0.171066   0.045922  -3.725 0.000227 ***
ahumid            0.006433   0.002832   2.271 0.023731 *
tsunamhour        0.130158   0.049866   2.610 0.009430 **
factor(holiday)1 -0.171998   0.053252  -3.230 0.001353 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4684 on 358 degrees of freedom
Multiple R-squared:  0.7024, Adjusted R-squared:  0.6974
F-statistic: 140.8 on 6 and 358 DF,  p-value: < 2.2e-16
```

분석 결과, R^2는 0.7024로 로그변환 전의 모형보다 0.061 상승했음을 확인할 수 있다. 또한 awind, ahumid, tsunamhour, holiday 변수들의 t값의 절대값이 로그변환 전의 모형보다 상승하여 각 변수의 설명력이 높아졌음을 확인할 수 있다.

로그변환한 회귀모형이 이분산성이 존재하는지를 확인하기 위해 White test를 실시한 결과, p값이 0.00007076으로 조금 개선되었으나 여전히 이분산성이 존재함을 확인할 수 있었다.

```
 studentized Breusch-Pagan test

data:  l
BP = 28.653, df = 6, p-value = 7.076e-05
```

(1)-2-2 FGLS (Feasible GLS)

GLS에 의해 구해지는 추정량은 반드시 가장 효율적인 불편향 선형추정량(BLUE; Best Linear Unbiased Esimator)이 된다. 하지만 우리는 각각의 σ_i를 모르기 때문에 이를 추정해야 한다. 즉, 오차의 분산에 영향을 미치는 요인의 추정치 $\hat{\sigma_i}$를 찾아 각각의 변수를 나눈 후 회귀분석을 하는 것이다. 여기서는 잔차의 절대값을 σ_i의 추정치로 사용했다.

σ_i의 추정의 어려움은 FGLS가 갖는 태생적인 한계이다. σ_i가 잘못 추정되었을 경우, 전체 회귀모형에 대한 편향(bias)이 발생할 가능성이 존재한다. 하지만 FGLS는 White's correction보다 보다 적은 수의 샘플에서 유리하며, 효율성 문제와 t검정의 문제를 둘 다 해결할 수 있기 때문에 다음과 같은 방법을 채택함을 명시한다.

```
Call:
lm(formula = lncount ~ mtemp + rtime + awind + ahumid + tsunamhour +
    factor(holiday), data = assemble, weights = 1/abs(residuals(l)))

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-1.1311 -0.6013  0.1376  0.6105  1.0177

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       9.039027   0.094690  95.459  < 2e-16 ***
mtemp             0.042800   0.002055  20.826  < 2e-16 ***
rtime            -0.070771   0.003980 -17.783  < 2e-16 ***
awind            -0.184553   0.020205  -9.134  < 2e-16 ***
ahumid            0.006778   0.001221   5.552 5.50e-08 ***
tsunamhour        0.140435   0.027356   5.134 4.68e-07 ***
factor(holiday)1 -0.161517   0.023910  -6.755 5.79e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6298 on 358 degrees of freedom
Multiple R-squared:  0.8761, Adjusted R-squared:  0.8741
F-statistic: 422.1 on 6 and 358 DF,  p-value: < 2.2e-16
```

분석 결과, R^2가 0.8761로 크게 상승하였음을 확인할 수 있다. 또한 모든 설명변수의 t값의 절대값이 유의미하게 상승하여 각 변수의 설명력이 크게 증가하였다. 특히 기본모형에서 낮은 설명력을 보였던 ahumid 변수 또한 FGLS 후에는 종속변수에 유의미한 영향을 미친다는 사실을 확인했다.

```
 studentized Breusch-Pagan test

data:  fgls
BP = 5.5388, df = 6, p-value = 0.4768
```

마지막으로, FGLS모형에 대한 이분산성 검정을 실시했다. White test의 검정통계량이 눈에 띄게

감소하여 p값이 0.4768으로 도출되었다. 따라서, 위의 FGLS모형에는 이분산성이 존재한다고 보기 어렵다.

(2) 자기상관

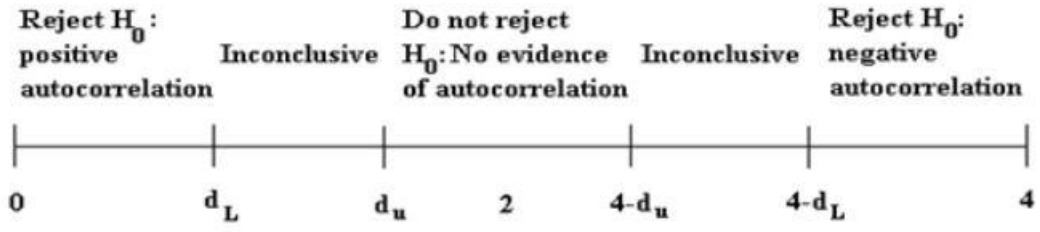
(2)-1 자기상관의 진단

시간 또는 공간적으로 연속된 일련의 관측치들 간에 존재하는 상관관계로서, 보통 시계열 데이터에 내재하는 시점 간의 상관이다. 시계열자료에서는 현재의 상태가 과거와 미래의 상태에 밀접한 연관을 지니는 경우가 많은데 이러한 경우를 자기상관이라고 한다.

이를 진단하기 위하여 Durbin Watson Test를 진행하였다. DW검정법이란 오차항에 자기상관이 있는지 없는지를 판단하기 위해 사용하는 검정법으로 오차항이 AR(1)을 따르고 모형에 설명변수로 lagged variable이 존재하지 않을 때 사용가능한 방법이다. 우리는 설명변수에 lagged variable을 넣지 않았고, 오차항이 지난 한 기수(즉 1일) 이전의 오차로부터만 영향을 받는다는 AR(1) 가정하에 이를 실시하기로 하였다. DW 통계량을 식으로 나타내면 아래와 같다.

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \fallingdotseq 2(1-\hat{\rho})$$

이 d통계량을 판단하는 방법은 아래와 같다.



DW검정법을 진행한 결과는 아래와 같았고 그 결과 DW 통계량은 0.48526으로 2보다 작기 때문에 양의 자기상관이 존재한다는 것을 알 수 있다.

```
> dwtest(fit3)

        Durbin-Watson test

data:  fit3
DW = 0.48526, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

(2)-2 자기상관의 해결.

자기상관을 해결하기 위해서 Cochrane-Orcutt 방법을 사용하였다. 이는 OLS가 같는 추정방법의 한계성을 보완하고, praise-winsten추정방법이 갖는 추론상의 우수성을 보유하는 방법이다.

```
> fit3_coc = cochrane.orcutt(fit3)
> dwtest(fit3_coc)

        Durbin-Watson test

data:  fit3_coc
DW = 2.6357, p-value = 1
alternative hypothesis: true autocorrelation is greater than 0
```

위와 같이 Cochrane-Orcutt 방법을 사용한 결과 DW 통계량이 2.637이 되어 자기상관 문제를 해결할 수 있었다. (p-value > 0.05)

```
> summary(fit3_coc)
Call:
lm(formula = log(count) ~ mtemp + rtime + awind + ahumid + tsunamhour +
    factor(holiday), data = data)

                   Estimate Std. Error t value  Pr(>|t|)
(Intercept)       9.3079269  0.1694354  54.935 < 2.2e-16 ***
mtemp             0.0339038  0.0049950   6.788 4.765e-11 ***
rtime            -0.0782351  0.0049394 -15.839 < 2.2e-16 ***
awind            -0.0141278  0.0260899  -0.542   0.58850
ahumid            0.0028730  0.0017725   1.621   0.10593
tsunamhour        0.0673849  0.0295633   2.279   0.02324 *
factor(holiday)1 -0.1456915  0.0280506  -5.194 3.468e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2821 on 357 degrees of freedom
Multiple R-squared:  0.7274 ,  Adjusted R-squared:  0.7228
F-statistic: 158.8 on 6 and 357 DF,  p-value: < 1.499e-97

Durbin-Watson statistic
(original):    0.48526 , p-value: 8.534e-49
(transformed): 2.63568 , p-value: 1e+00
```

자기상관을 없애고 다중회귀모형에 다시 대입시킨 결과 R-square 값이 상승하였다. 그러나 두 개의 독립변수(awind, ahumid) 가 유의미하지 않다고 나오는 한계가 존재한다.

(3) 다중공선성

(3)-1 다중공선성의 진단

다중공선성의 문제는 단순회귀분석에서는 발생하지 않고 다중회귀분석에서 발생한다. 다중공선성은 설명변수 간 상호의존성이 매우 높은 경우, 또는 상관계수가 매우 높은 경우에 발생한다. 만약, 2개의 독립변수가 서로 상관관계가 높을 경우, 어느 하나의 독립변수가 종속변수의 변동을 설명하고 다른 변수가 남은 변동을 설명하게 되므로 설명할 수 있는 변동은 작아지고 표준오차는 증가하게 된다. 따라서 회귀모형의 유의성을 검정하는 t-통계량은 표준오차가 분모이므로 t-통계량이 작아져서 두번째 변수는 유의하지 않게 되는 사례가 발생한다. 즉, 단순회귀에서는 유의하다가 다중회귀에서는 유의하지 않게 되는 모순되는 결과가 발생하는 것이다.

다중공선성의 존재 여부를 진단하기 위해 위의 모형의 분산 팽창 인수(VIF)를 조사하였다. VIF는 예측변수가 상관 관계가 있는 경우, 추정된 회귀 계수의 분산이 증가하는 정도를 측정하는데 일반적으로 10을 넘으면 다중공선성이 존재한다고 판단한다.

```
> vif(fit3)
     mtemp      rtime      awind     ahumid tsunamhour factor(holiday)
  3.133232   2.589000   1.307048   3.021649   3.089155       1.022662
```

그 결과는 위와 같다. 다중공선성은 검정의 문제가 아니라 정도의 문제이기 때문에 연구자가 기준을 정해 판단할 수밖에 없다. 우리는 10을 기준으로 설정했고 따라서 모형의 설명변수간 다중공선성이 존재하지 않는다고 판단할 수 있다.

(4) 내생성

우리는 내생성에 대한 분석은 진행하지 않았는데 이는 날씨와 휴일이 선험적으로 결정된 요소이고 종속변수가 독립변수에 영향을 끼치는 현상인 역인과관계가 따릉이 이용건수와 날씨 및 휴일 사이에서는 있을 수 없기 때문이다. 따릉이 이용횟수가 자연이 결정하는 요소인 날씨와 미리 지정되어 있는 휴일 및 공휴일 여부에 영향을 미칠 수 없다는 것은 자명한 사실이다. 물론 나중에 따릉이 뿐만 아니라 공유 스마트모빌리티 개념이 더욱 보편화되어 환경오염 및 대기오염이 감소하고 저탄소 사회가 되어서 지구온난화 등에 끼치는 영향이 감소, 결국 기온 등에 영향을 미치는 등의 인과관계를 가질 수는 있으나 이는 수십년에 걸쳐 일어날 변화이고 '일별' 사용량이 탄소절감 및 기온 변화에 미치는 영향은 미미할 것이기 때문에 내생성을 고려하지 않기로 하였다.

3. 분석 결과

기본모형의 회귀분석에서는 이분산성과 자기상관성이 나타나 신뢰할만한 분석결과를 얻을 수

없었다. 따라서 종속변수의 로그변환과 FGLS를 통해 모형의 이분산성 문제를 해결했고 자기상관성 문제는 Cochrane-Orcutt 방법으로 해결했다.

그 중 로그 변환을 통해 이분산성 문제를 해결한 회귀식을 제시하자면 다음과 같다.

$$\hat{Y}_{lncount} = 9.039027 + 0.0428X_{mtemp} - 0.069313X_{rtime} - 0.171066X_{awind} + 0.006433X_{ahumid} + 0.130158X_{tsunamhour} - 0.171998X_{holiday}$$

회귀분석에서 로그 변환을 했을 경우 계수 해석에 유의해야 한다. 종속변수 Y에 로그가 취해진 경우에 각 계수의 추정치는 X의 한 단위 변화에 대한 Y의 변화율을 의미한다.

위 회귀식의 R^2는 0.8761로 기본모형의 R^2보다 크게 상승하였음을 확인할 수 있다. 또한 모든 설명변수들의 p값이 충분히 낮게 나와 모형이 적절함을 확인할 수 있다. 먼저 종속변수인 따릉이 이용횟수와 양의 관계를 갖는 설명변수에는 최고기온(mtemp), 평균상대습도(ahumid), 1시간 최다일사량(tsunamhour)이 있다. 최고기온이 섭씨 1도 상승할 때마다 따릉이 이용횟수는 4.28% 증가함을 확인할 수 있다. 또한 평균상대습도가 1% 상승할 때마다 따릉이 이용횟수는 약0.64% 증가하고, 1시간 최다일사량이 한 단위 상승할 때는 약 13%가 증가한다.

반대로 따릉이 이용횟수와 음의 관계를 갖는 설명변수에는 계속강수시간(rtime), 평균풍속(awind), 휴일(holiday)이 있다. 계속강수시간이 1시간 늘어날 때마다 따릉이 이용횟수는 약 6.93% 감소하며, 평균풍속이 한 단위 늘어날 때마다 따릉이 이용횟수는 17.1% 정도 감소한다. 마지막으로 휴일 설명변수에 대한 해석은 로그 변환 시 가변수에 대한 해석이기 때문에 계수 그대로 해석해서는 안된다. $e^{0.171998} = 1.187675\ldots$ 이므로, 휴일일 때 18.76% 정도 따릉이 이용횟수가 감소함을 확인할 수 있다.

3. 결론

1. 요약

우리는 환경 오염에 대한 경각심과, 공유 경제 중 특히 스마트 모빌리티 사업에 대한 관심이 점차 상승하고 있는 현 상황 속에서 따릉이의 이용량이 중요한 시사점을 지닌다고 보고 이에 대한 분석을 기상요인을 중심으로 진행하였다. 분석 결과 각 가정 별 solution을 진행하기 이전과 진행한 이후 모두, 2018년 따릉이 총 이용횟수는 "최고기온(mtemp), 평균상대습도(ahumid), 1시간 최다일사량(tsunamhour)"과는 양의 관계를, "계속강수시간(rtime), 평균풍속(awind), 휴일(holiday)"과는 음의 관계를 지님을 확인할 수 있었다. 이를 통해 따릉이의 이용량이 온도가 낮지 않고, 해가 많이 비추고, 강수가 적고 바람이 많이 불지 않을 때 이른바 자전거를 타기 좋은 날씨가 갖춰질 때 늘어난다는 점을 관찰할 수 있었으며 이는 우리의 분석 이전의 생각과도 일치하였다. 반면 평균상대습도(ahumid)와 휴일(holiday)에 있어서는 분석을 진행하기 전의 추론과 반대 결과가 나왔다. 분석 이전에는 습도가 높아지면 따릉이 이용량이 감소할 것이고 휴일에 오히려 따릉이 이용량이 증가할 것이라고 추측하였기 때문이다. 이를 통해 습도가 오르면 사람들의 불쾌함이 커지고 걷기 보다는 자전거를 타는 것을 택하기 때문에 따릉이 이용량이 증가하고 따릉이의 이용량이 여름에 집중되어 있는데 여름에는 습도가 높기 때문에 이와 같은 결과가 나온 것이라고 해석할 수 있었다. 또한 따릉이의 이용 목적이 휴일의 여가용 보다는 출퇴근시 이용인 사람들의 수가 많아 평일에 이용량이 증가한 것이라는 추측을 가능하게 해주었다.

2. 한계 및 제언

먼저 우리는 우리의 기존 모형에서 발생한 자기상관과 이분산성을 동시에 해결하는 모형을 제시하지 못하고 각 해결 과정을 개별적으로 제시할 수밖에 없었다. 또한 데이터의 부족으로 인해 따릉이의 지역별 이용량 분석이 아닌 총 이용량 분석을 진행할 수밖에 없었다.

만일 지역별 분석, 특히 출퇴근시 유동인구가 많은 지역과 한강 공원과 같이 휴일에 유동인구가 많을 것으로 예상되는 지역을 구분하여 분석을 진행할 수 있었다면 휴일 여부에 따른 이용량 및 기타 기상요인이 미치는 영향에 관해 좀 더 정확한 분석을 진행할 수 있었을 것이다. 또한 만일 데이터를 구할 수 있어 지역별 분석을 진행하였다면, 각 지역별로 이용에 영향을 미치는 다른 변수들을 추가하여 우리의 분석과 조금은 다른 결과를 얻어낼 수도 있었을 것이다. 예를 들어, 따릉이를 주로 출퇴근시 자가용이 아닌 지하철 같은 대중교통을 이용하는 사람들이 탄다는 합리적인 가정을 추가한다면 각 지역별(대여소별) 지하철과의 거리가 이용량에 미치는 영향 등을 추가적으로 분석할 수 있었을 것이기 때문이다. 이후에 지역별 자료 데이터가 충분히 존재하고 서론에서 밝혔던 데이터를 구할 수 없어 제외했던 기타 변수들의 자료를 구할 수 있다면 따릉이 이용량

에 대해 새로운 분석을 진행할 수 있으리라고 생각한다.

3. 시사점

따릉이 일별 총 이용횟수에 영향을 미치는 날씨에 대한 분석은 앞으로 스마트 모빌리티 사업이 더욱 확장됨에 따라 보다 큰 의미를 가질 수 있다. 앞으로는 거치대 및 대여소가 존재하지 않고 길거리에 자유롭게 스마트 모빌리티가 나와있게 될 것이며 서울시도 실제로 올해 11월에 대여소가 없는 전기 자전거를 도입할 것을 계획하고 있다. 대여소와 그 대여소에 놓인 거치대 수가 정해져 있어 날씨에 상관없이 항상 지정된 장소를 지켜야 하는 기존의 방식과 달리, 앞으로는 날씨에 따라 그 위치를 바꿀 수 있다는 장점을 살려 날씨 변수를 고려하여 사람들의 이용을 더욱 활성화할 수 있을 것이기 때문이다. 비록 우리는 서울시 전체의 총 이용량에 대해 날씨에 따른 분석을 진행하였지만, 이후 일별로 대여소가 없는 전기 자전거를 배치할 때, 서울 내의 각 구역의 날씨를 미리 기상청과의 협조를 통해 살펴보고, 이용량이 클 것으로 생각되는 곳에 사전에 모빌리티를 많이 배치두는 데에 활용될 수 있을 것이다. 만일 서울 시내에서는 지역별 날씨가 크게 차이가 발생하지 않을 것이라고 생각한다면, 위와 같은 스마트모빌리티 사업이 전국 수준으로 확장되었을 때 지역별로 미리 스마트 모빌리티를 배치하는 데에 활용하는 것으로 나아갈 수 있을 것이다. 앞으로 이용량이 많은 곳에 따릉이 및 스마트모빌리티를 배치함으로써 사람들의 이용을 더욱 활성화하고, 이러한 공유경제가 더욱 성장해 나간다면 더 깨끗한 지구를 기대해볼 수 있을 것이라고 생각한다.

<참고 자료>

기상청 관측자료. www.kma.go.kr

서울 시설공단. "공공자전거 운영처_종합현황(2019.04)" http://new.sisul.or.kr/open_content/main/

서울 열린 데이터 광장, "서울특별시 공공자전거 이용정보(일별)". http://data.seoul.go.kr/