회귀분석Ⅰ 학기말 과제 최종보고서

다중 선형 회귀모형을 이용한 한남대교의 교통량 분석

서울시립대학교 통계학과

2017580024

임성준

목 차

1. 서 론 ····· 1

1.1 연구 목적, 주제 및 개요 ····· 1

1.2 문헌 연구 ····· 1

1.3 데이터 설명 ····· 2

1.4 분석 방법 : 회귀분석 I ····· 3

1.5 결과 활용 및 기대 효과 ····· 4

2. 본론 ····· 5

2.1 탐색적 자료 분석 (EDA) ····· 5

2.2 모형 적합 (Model Fitting) ····· 10

2.3 모형 타당성 점검 ····· 17

3. 결론 ····· 18

3.1 분석 결과 요약 ····· 18

3.2 분석의 장점 및 한계점 설명 ····· 20

3.3 추가 연구사항 제안 ····· 20

4. 참고문헌 ····· 21

1. 서론

1.1. 연구 목적, 주제 및 개요

이번 연구를 통해 우선 현존하는 총 31개의 한강 대교 중 한남대교를 바탕으로 교통량에 영향을 줄 수 있는 요인들을 통계적으로 파악하고자 한다.

우선적으로 한강 교량 교통량에 영향을 줄 수 있는 요인으로 총 4개의 후보군을 선정하였다. ① 코로나 확진자 수, ② 기상, ③ 연료 가격, ④ 지하철 이용자 수이 이에 해당한다. 앞의 4가지 요인을 설명변수, 각각의 교량의 교통량을 반응변수로 지정하여 다중 선형 회귀 모형을 바탕으로 한 회귀분석을 실시해보았다.

과정은 크게 다음과 같다. 우선 수집한 데이터를 사용에 용이하도록 전처리를 가해주었다. 그 후, 데이터를 우선적으로 상관분석과 그래프와 그림을 통한 직관적인 검토를 하는 탐색적 자료분석을 실시하였다. 그 다음 단계는 4가지 요인을 한남대교의 교통량에 대해 직접 모형에 적합해보는 과정을 실시했다. 여기서 통계적으로 유의하지 않은 변수들을 제거해 나가면서 다양한 모형에 대한 적합을 시도했다. 마지막으로는 적합한 모형이 회귀분석을 위한 기본 가정을 만족시키는지 확인해보았다. 이 때 잔차분석을 통해 이를 실행했다. 마지막으로 적합된 모형을 바탕으로 한남대교에 대한 교통량을 예측해보는 시간을 가졌다.

1.2. 문헌 연구

진행하고자 하는 연구와 유사한 연구로는 서울시 교통정보 시스템(이하 TOPIS)에서 매년 발표하는 '서울시 교통량조사 보고서'가 존재한다. 이는 서울 전역의 교통량을 총 135개 지점을 통해 조사하는 자료이다. 이 중에는 현재 연구 대상으로 선정한 한남대교를 포함한 20개의 교량 지점이 포함된다. 이 보고서에서는 전년대비 교통량 현황을 분석하며, 지점별 월별 요일별 교통량과 지점별 월별 시간대별 교통량, 월별·요일별·시간대별 교통량 변동계수를 제공한다.

TOPIS에서는 또한 월 단위로 지점별 일자별 교통량을 제공하고 있다. 이 조사에서 또한 20개 교량의 교통량에 대한 조사 역시 진행되었다. 이는 단순한 조사 자료로, 어떠한 요인들이 교통량에 영향을 미치는지에 대한 연구는 진행되지 않았다는 점에서 진행하고자 하는 연구와는 차이가 있었다.

한편, 『대도시의 대중교통수요 영향요인 분석』와 같이 대중교통에 대한 영향요인 분석은 존재했지만 특정 도로를 중심으로, 대중교통만을 고려한 것이 아니라 전체적인 교통량을 고려하여 분석을 진행한 연구는 발견하지 못하였다.

1.3. 데이터 설명

이 연구에서 설명변수, 반응변수를 포함해 사용되는 모든 변수는 총 6개이며, 총 11개의 데이터셋이 사용된다. 먼저, 반응변수로는 교통량에 대한 데이터로 TOPIS에서 제공하는 월별 서울시 교통량 조사 자료를 이용했다. xlsx 확장자로 제공이 되었지만 SAS에서 사용하기 쉽도록 csv 확장자로 저장해 사용했다. 현재 관심 있는 한남대교에 대한 자료만 이용했으며 유입과 유출의 평균을 이용해 1월 1일부터 3월 31일의 한남대교 교통량을 추정해 traffic이라는 이름의 데이터셋으로 저장하였으며 각각을 회귀분석에 이용하였다.

두 번째로 이용한 자료는 코로나 확진자 수 데이터이다. 이는 한국데이터거래소(KDX)를 통해 무료로 접근할 수 있었다. 주어진 csv 파일에서 EXCEL의 간단한 함수를 이용해 누적 확진자수를 바탕으로 일일 확진자 수를 계산해 하나의 설명변수로 이용했다. 이렇게 처리한 데이터를 covid_19이라는 데이터셋으로 저장했다.

세 번째로 이용할 자료는 기상 관련 데이터이다. 이 자료는 기상자료개방포털에서 다운로드했다. 기상자료개방포털에서 다양한 자료를 제공하지만 그 중에서 일별 강수 계속시간, 일강수량, 일 최심적설[1], 일 최심적설 시간, 평균 지면온도, 안개 계속시간 관련 정보를 선택했다. 이 중, ① 강수량, ② 최심적설, ③ 안개 계속시간을 주목하려 했으나 1월부터 3월까지 안개가 거의 존재하지 않아 최종적으로 강수량, 최심적설만 이용하기로 하였다. 강수량이 5mm 이상인 경우 1, 그렇지 않은 경우 0으로, 최심적설이 존재하면 1, 존재하지 않으면 0에 해당하는 더미 변수로 만들어 회귀 분석에 사용 가능케 하였다. 따라서, 기상 관련 2개의 설명변수가 추가된다.

네 번째로 이용하는 자료는 연료가격에 관한 데이터이다. 이 데이터의 출처는 Opinet이라는 웹사이트이며, 이 또한 무료로 사용 가능하였다. 여기서는 일자별 전국 주유소의 고급휘발유, 보통휘발유, 자동차용경유의 평균을 제공해준다. 고급휘발유와 보통휘발유의 평균을 휘발유 가격으로 사용하며, 서울시 자동차 등록 현황을 바탕으로 휘발유 차량과 경유 차량의 비율을 구해 휘발유와 경유 가격을 가중 평균해 연료 가격이라는 하나의 설명변수로 이용한다. 이 때 서울시 자동차 등록 현황은 서울 열린 데이터 광장을 통해 이용하였고, 이는 아래의 <표1-1>과 같다. 1월부터 3월까지 휘발유와 경유 차량의 총합 (A+B+C+D+E+F)은 8,225,868이고 1월부터 3월까지의 휘발유 차량의 총합(A+C+E)은 4,908,823이며, 동일 기간 경유 차량의 총합(B+D+F)은 3,317,045이다. 이를 통해 2021년 1분기 휘발유 차량과 경유 차량의 비율은 6:4임을 계산해낼 수 있다.

1) 최심적설: 고려하고 있는 기간 동안, 전부터 내려 녹지 않고 쌓여 있을 눈을 포함하여 가장 두껍게 쌓여 있을 때의 눈의 두께(깊이).

기간	차종별	계	휘발유	경유	LPG	전기	CNG	하이브리드	수소
2021.01	계	3,159,552	1,636,281 (A)	1,107,835 (B)	257,262	23,441	9,112	117,578	1,719
2021.02	계	3,158,102	1,635,844 (C)	1,106,027 (D)	256,291	23,381	9,100	119,382	1,698
2021.03	계	3,158,710	1,636,698 (E)	1,103,183 (F)	255,058	24,918	9,073	121,605	1,697

<표1-1> 서울시 자동차 등록 현황 (2021년 1월~3월)

마지막 자료는 대중교통 데이터이다. 대중교통 중 교량의 교통량에 직접적으로 집계되지 않는 지하철만을 대상으로 이용하였다. 이 자료는 서울 열린 데이터 광장을 통해 이용하였으며 2021년 1월부터 3월까지 각 지하철 역별 승하차 인원 정보를 제공 받았다. 일자별로 평균 지하철 이용객 수를 계산하여 하나의 설명변수로 사용하였다. 이 때, 승차 총 승객과 하차 총 승객의 평균을 한 지하철역의 승객으로 간주하였다.

이렇게 고려한 설명변수들을 한남대교의 교통량과 연결하여 traffic 데이터셋으로 종합한다. 따라서 하나의 데이터셋을 이용해 회귀분석을 실시할 수 있도록 한다.

1.4. 분석 방법

본 연구를 진행하기 위해 기본적으로 다중 선형 회귀 모형을 이용한다. 다중 회귀 모형은 단순 회귀 모형이 확장된 모형으로 하나의 반응변수에 두 개 이상의 설명변수가 사용된다. 따라서 총 5개의 설명변수가 사용되는 본 연구에 적용 가능하다고 할 수 있다. 한편, 다중 선형 회귀 모형을 받아들이기 위해선 먼저 회귀 모형을 위한 가정이 만족되어야 한다. 등분산성 가정, 선형성 가정 등이 존재하며 가정이 만족되지 않는다면 아무리 모형 설명력이 좋은 모델로 적합이 된다고 해도 그 모형은 사용할 수 없다. 따라서 가정이 충분히 만족되는지 확인하는 과정이 반드시 필요하다. 이를 확인하기 위해 residual plot, normal probability plot 등을 이용한 잔차분석을 진행한다. 만약 가정에 대한 위반이 확인된다면 가정을 만족시키기 위해서 여러 가지 model transformation을 진행해봐야 한다.

다중 선형 회귀 모형의 가정 충족 여부에 관계없이 데이터가 모형에 적합될 때 least square method를 통해 intercept를 비롯한 각 coefficient의 추정치가 계산된다. 이렇게 계산된 추정치가 실제로 유의한지, 즉 0이 아닌지 확인하기 위해 H_0 : $\beta_i = 0$, i=1, 2, 3, 4, 5에 대해 가설 검정을 실시한다. 가설 검정을 통해 유의확률(p-value)이 유의수준 5%보다 작으면 H_0를 기각한다. 즉, $\beta_i \neq 0$이며 추정된 회귀계수가 유의미하다고 판단한다. 다음의 가설 검정을 모든 설명변수에 걸쳐 진행하고 H_0를 기각하지 못하는, 즉 $\beta_i = 0$인 회귀계수에 해당하는 변수를 제거하고 다시 회귀분석을 실시한다. 또한 여러 변수 조합을 시도해보면서 최상의 R-Square 값이 나오는 모델을 찾는다. 여기서 다중 회귀 분석임에 주의해서 Adjusted R-Square 값을 참고한다.

또한, 모든 관측치가 하나의 모델에 동일한 영향을 끼치지 않음에 유의하여 influence point를 탐색하는 과정을 거쳐야한다. 이 과정은 leverage point와 influence point가 존재하는지 탐색하는데 적합된 결과에 큰 영향을 끼쳐 왜곡된 결과를 이끌어내는지 주목해야 한다.

다중 선형 회귀 모형을 적합하면서 다중공선성에 대해서도 신경 써야 한다. 특히, 이번 조사에서는 다중공선성이 존재할 것이라고 예상되는 설명변수들이 존재한다. 따라서 이에 대해 의심되는 변수들을 따로 산점도를 그려 둘 간의 선형 관계가 존재하는 확인하고, 추가로 VIF, correlation matrix, condition number를 이용해 다중공선성의 유무를 파악하는 단계가 꼭 필요하다.

즉, 우선적으로 다양한 모델을 적합해보며 모형 설명력이 좋은 모형을 찾고, 그 모형이 선형 회귀 모형의 가정을 충족하는지, outlier와 같은 influence point가 존재하는지 확인하는 과정을 반복해야 한다. 또한 설명변수들 간의 다중공선성의 유무를 파악하고 이를 제거하는 과정 또한 필수적이다. 일련의 과정을 통해 선형 회귀 모형의 가정을 충족시킴과 동시에 모형 설명력이 좋은 모형을 최종적으로 선택해야 한다.

1.5. 결과 활용 및 기대 효과

이 연구는 총 31개의 한강 대교 중 한남대교만을 대상으로 교통량에 미치는 요인을 조사해보기 위해 진행되었다. 조사된 요인들을 바탕으로 교통량 증가/감소 요인을 분석함으로서 한남대교의 교통 체증을 감소할 수 있는 방안의 모색을 기대할 수 있다. 더 나아가 더 많은 수의 한강 대교를 같이 방법으로 분석하는 과정을 추가한다면 한강 대교들의 교통 체증을 해소할 새로운 방안을 제공할 수 있는 기초가 될 수 있을 것이다.

2. 본론

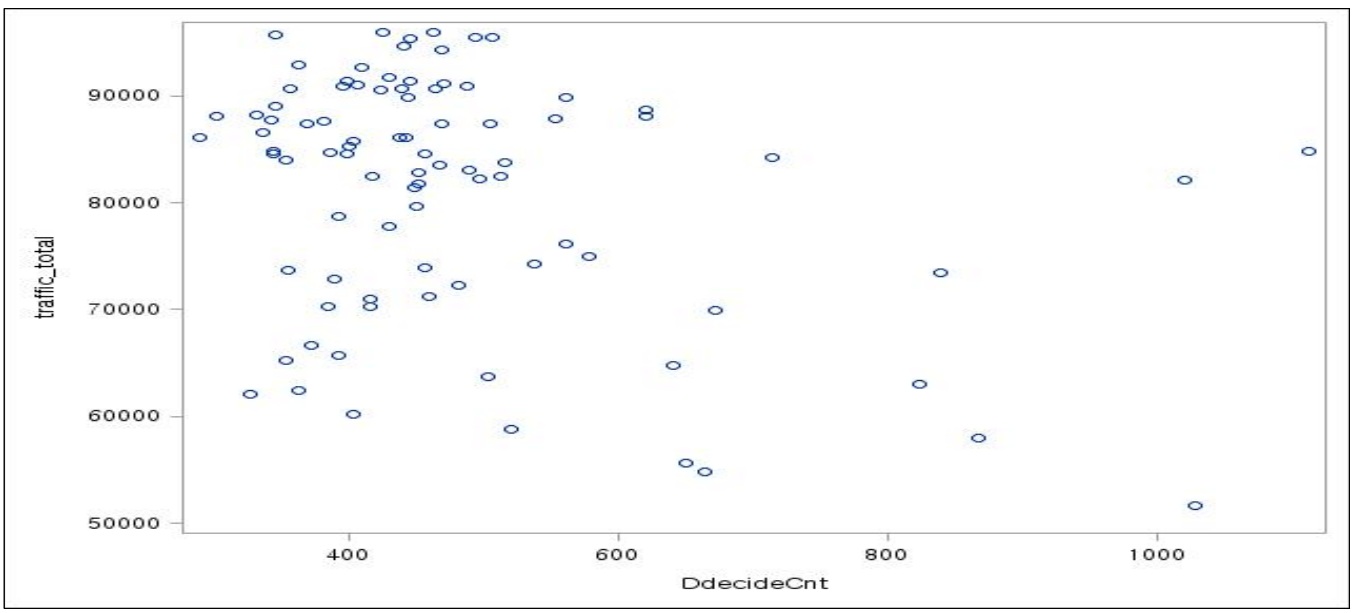
2.1. 탐색적 자료 분석(EDA)

회귀 모형에 적합하기에 앞서 사용할 데이터가 어떠한 정보를 담고 있는지 파악해볼 필요가 있다. 반응변수와 설명변수를 포함한 총 6개의 데이터 중 dummy_rain(비)과 dummy_snow(눈)만이 범주형 변수이며 나머지 4개 변수는 수치형 변수에 해당한다. 반응변수인 traffic_total을 포함한 4개의 수치형 변수에 대해서 요약통계량을 살펴보고, sgplot 프로시저를 이용해 산점도를 그려보았다. 더 나아가 반응변수인 traffic_total(한남대교의 교통량)과 설명변수인 Ddecidecnt(일일 코로나 확진자 수), users(지하철 이용자 수), fuel(연료 가격) 간의 상관계수에 대해 알아보았다.

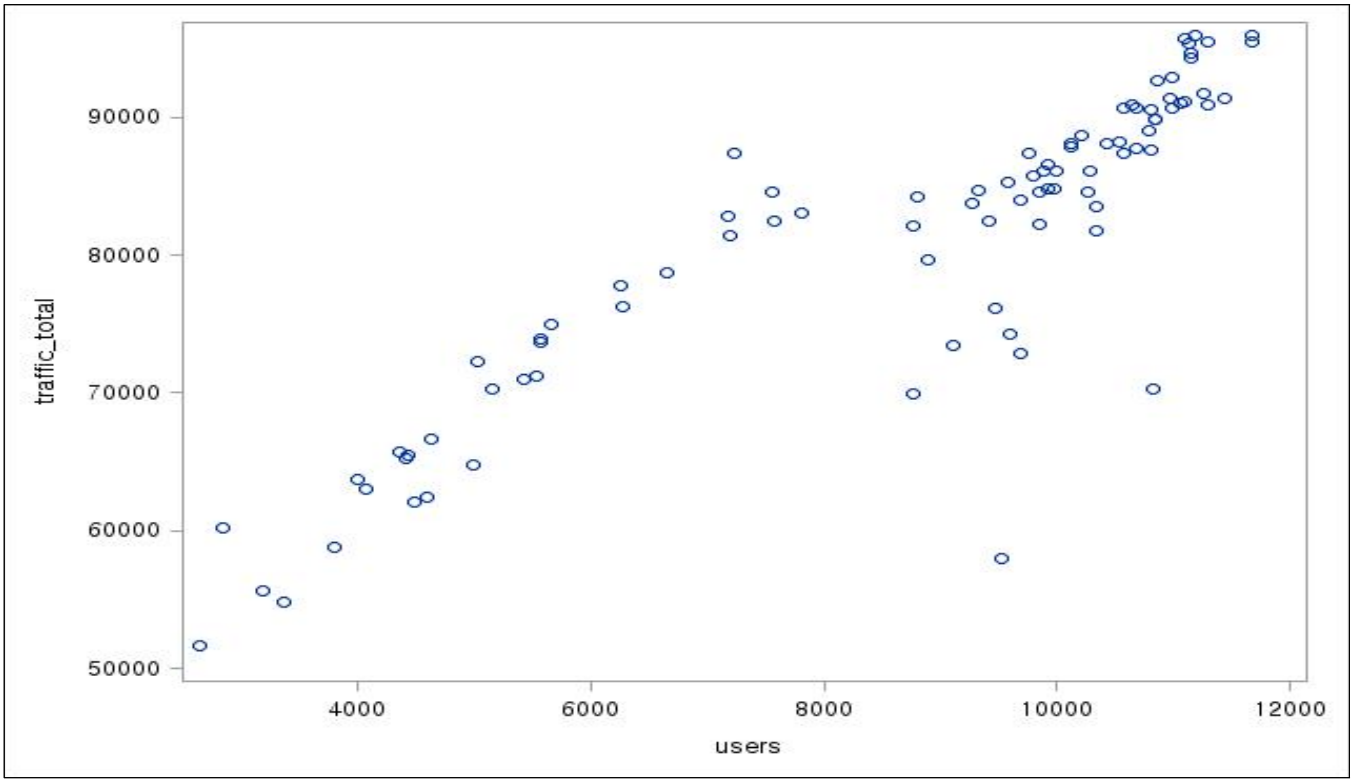
변수	N	평균	표준편차	최솟값	최댓값
traffic_total	90	80987.68	11191.87	51640.50	95938.50
DdecideCnt	88	481.2272727	156.5093899	289.0000000	1113.00
users	90	8634.88	2626.85	2633.67	11674.41
fuel	90	1464.61	32.8503146	1415.91	1523.92

<표2-1> 수치형 변수들의 요약 통계량

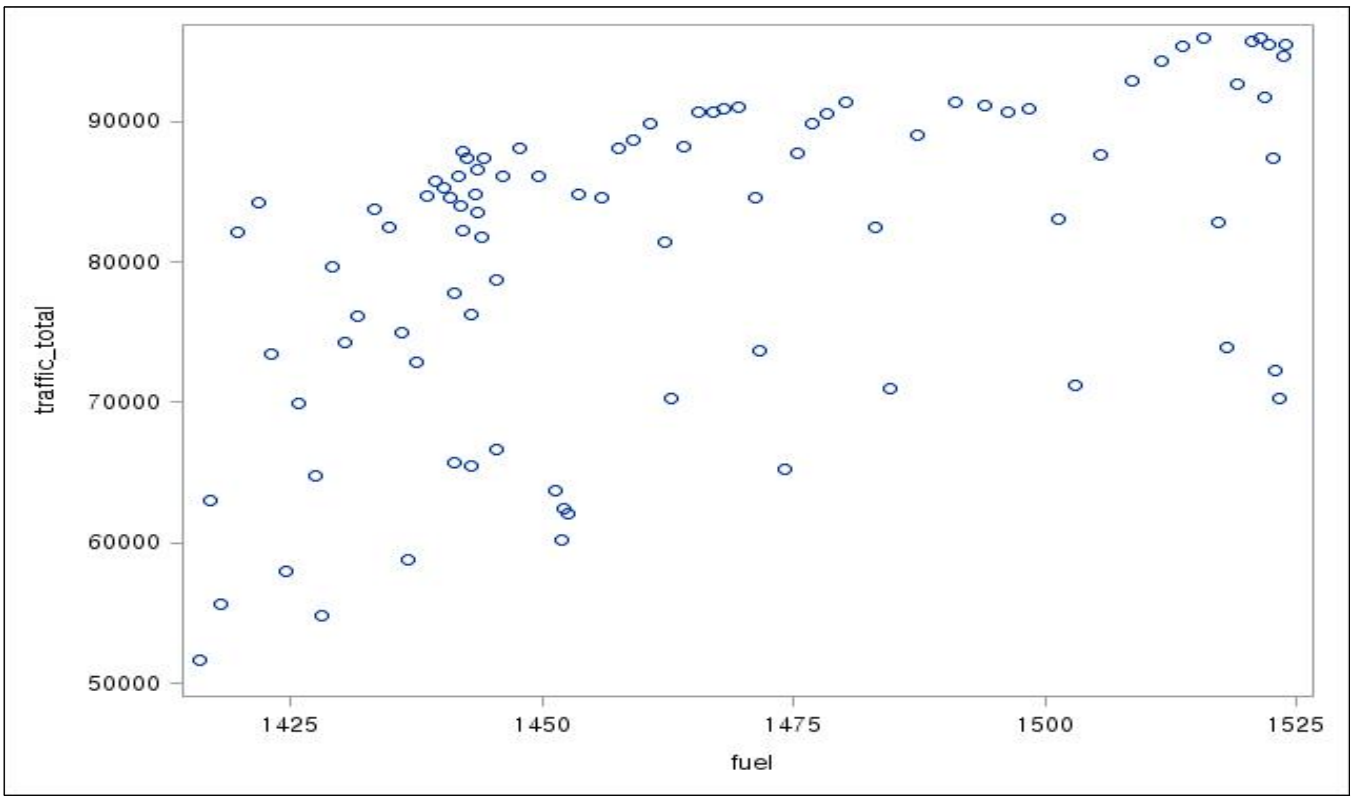
<표2-1>과 같이 수치형 변수들의 요약 통계량을 SAS의 means 프로시저를 이용해 구하였다. 여기서 DdecideCnt가 88개로 2개의 결측치를 포함하고 있는 것으로 나타났다. 이는 회귀분석을 실시할 때, 결측치가 포함된 행을 제거하는 방식으로 대응할 것이다. 또한 코로나 일일 확진자수의 최대값이 평균에 비해 매우 큰 형태를 띠고 있다는 것을 파악할 수 있다.



<그래프2-1> DdecideCnt와 traffic_total 간의 산점도



<그래프2-2> users와 traffic_total 간의 산점도



<그래프2-3> fuel와 traffic_total 간의 산점도

회귀분석을 실시하기 전에 반응변수 traffic_total과 각 수치형 설명변수들 간의 산점도를 그려보는 과정을 통해 대강 어떤 선형관계를 나타낼지 예상해보았다. 코로나 일일 확진자 수와 교통량 간에는 약한 음의 상관관계가 존재할 것으로 보인다. 반면, 지하철 이용자수와 교통량 사이에는 강한 양의 상관관계가 존재하는 것처럼 보인다. 마지막으로 연료가격과 교통량 사이에는 양의 상관관계가 존재할 것으로 예상되었다.

피어슨 상관 계수 H0: Rho=0 가정하에서 Prob > |r| 관측값 개수		
	traffic_total	DdecideCnt
traffic_total	1.00000 90	-0.32482 0.0020 88
DdecideCnt	-0.32482 0.0020 88	1.00000 88

<표2-2> DdecideCnt와 traffic_total 간의 산점도

피어슨 상관 계수, N = 90 H0: Rho=0 가정하에서 Prob > |r|		
	traffic_total	users
traffic_total	1.00000	0.87699 <.0001
users	0.87699 <.0001	1.00000

<표2-3> users와 traffic_total 간의 산점도

피어슨 상관 계수, N = 90 H0: Rho=0 가정하에서 Prob > |r|		
	traffic_total	fuel
traffic_total	1.00000	0.51253 <.0001
fuel	0.51253 <.0001	1.00000

<표2-4> fuel와 traffic_total 간의 산점도

<그래프2-1>, <그래프2-2>, <그래프2-3>에서 그림을 통해 예상해보았던 상관관계를 SAS의 corr 프로시저를 이용해서 직접 구했다. 그 결과 예상했던 것과 거의 일치하게 교통량과 코로나 일일 확진자수, 대중교통 이용자수, 연료가격은 각각 약한 음의 상관관계, 강한 양의 상관관계, 양의 상관관계를 나타냈다. 세 설명변수 반응변수에 대해 모두 유의미한 선형관계를 보이므로 교통량에 영향을 미치는 요인으로 간주할 수 있다.

한편 범주형 변수들을 파악해보기 위해 우선적으로 means 프로시저를 이용해 요약통계량을 살펴보았다. 그리고 boxplot 프로시저를 이용해 traffic_total에 dummy_rain과 dummy_snow 변수를 대응해서 boxplot을 그렸다. 이 때 boxplot 프로시저를 이용하기 위해 sort 프로시저를 이용해서 먼저 정렬을 해주었다.

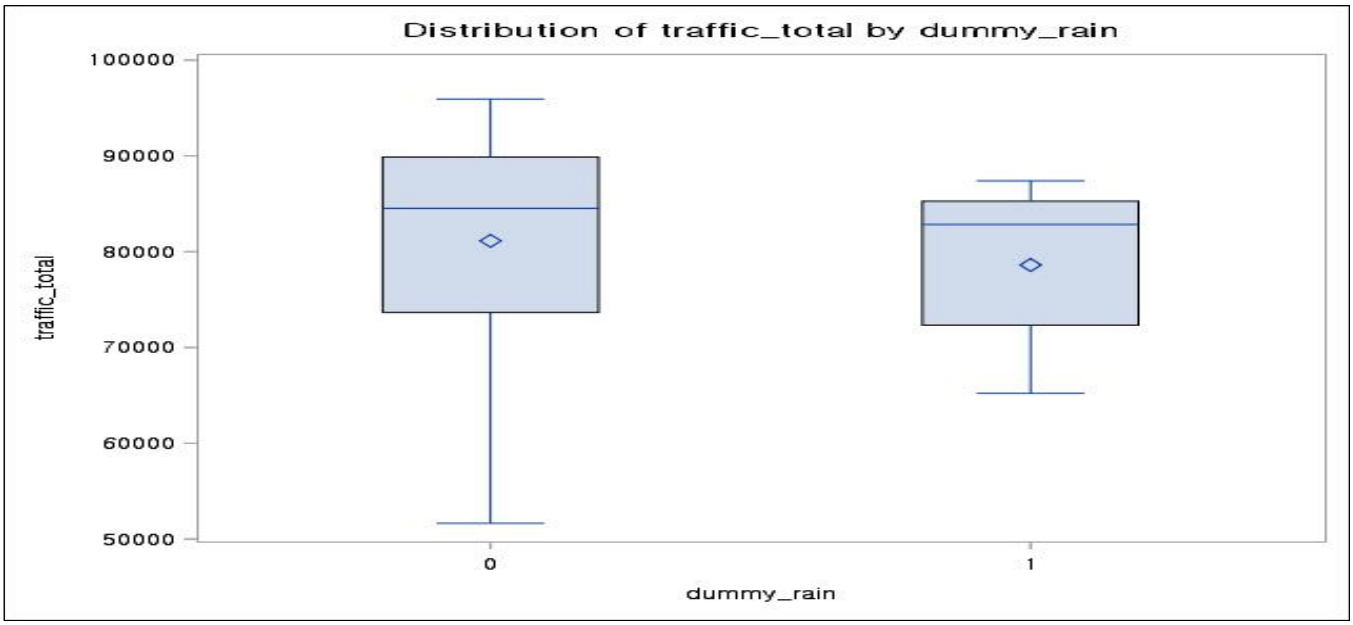
분석 변수 : traffic_total						
dummy_rain	관측값 수	N	평균	표준편차	최솟값	최댓값
0	85	85	81127.43	11317.55	51640.50	95938.50
1	5	5	78611.90	9470.78	65205.50	87400.00

<표2-5> dummy_rain에 따른 traffic_total의 요약 통계량

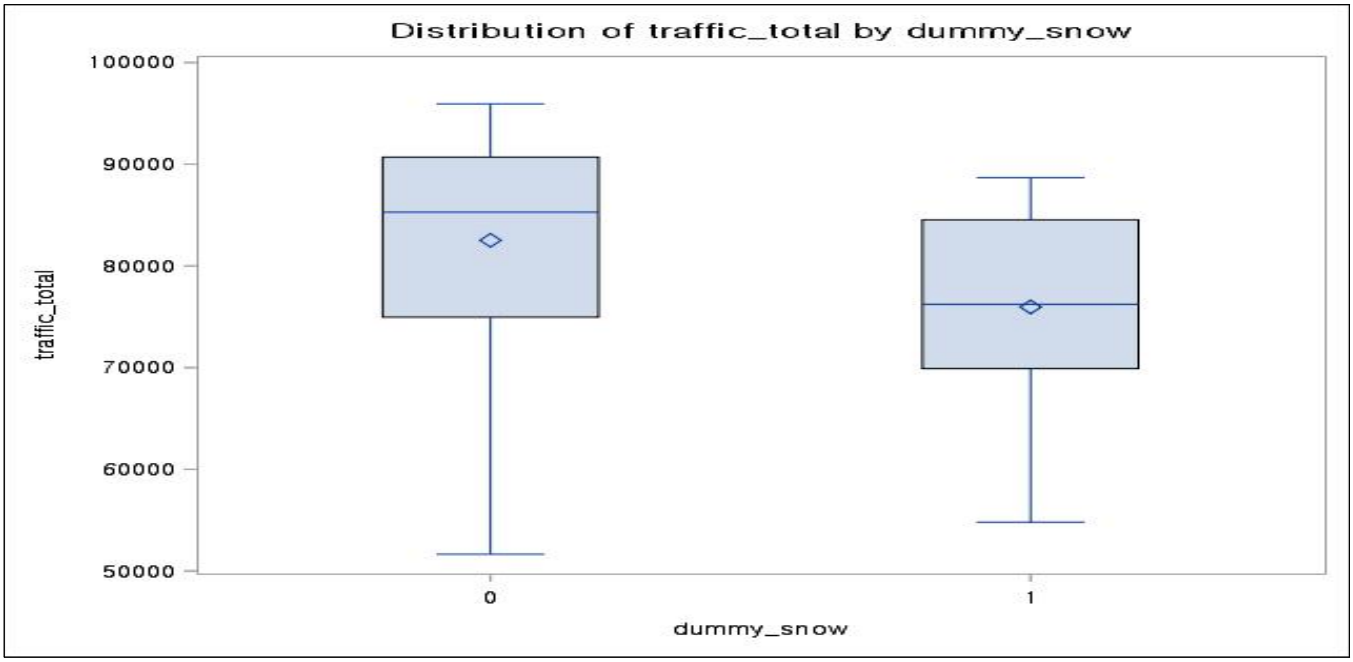
분석 변수 : traffic_total						
dummy_snow	관측값 수	N	평균	표준편차	최솟값	최댓값
0	69	69	82516.70	10963.63	51640.50	95938.50
1	21	21	75963.76	10683.94	54802.50	88687.50

<표2-6> dummy_snow에 따른 traffic_total의 요약 통계량

위의 <표2-5>와 <표2-6>과 같이 dummy_rain과 dummy_snow에 따른 traffic_total의 요약통계량을 구해보았다. 비가 온 날, 즉 dummy_rain이 1인 날, 교통량의 평균은 81,127.43에서 78,611.9으로 줄었고, 눈이 온 날, 즉 dummy_snow가 1인 날, 교통량의 평균은 82,516.70에서 75,963.76으로 줄었다는 점이 눈에 띈다.



<그래프2-4> dummy_rain과 traffic_total 간의 box-plot



<그래프2-5> dummy_snow과 traffic_total 간의 box-plot

위의 <그래프2-4>와 <그래프2-5>와 같이 dummy_rain, dummy_snow와 traffic_total 간의 box-plot을 그렸다. 앞선 요약통계량에서 살펴보았듯이 눈이 오거나 비가 오는 날에는 교통량이 감소하는 점을 파악할 수 있다. 따라서 한남대교 교통량에 영향을 끼치는 요인으로 간주할 수 있다.

2.2. 모형 적합

2.2.1. 첫 번째 모형

첫 번째로 고려한 모든 설명변수를 포함한 다중 회귀 분석을 실시해보았다. SAS의 reg 프로시저에서 반응변수로 한남대교의 교통량, 설명변수로는 일일 코로나 확진자 수, 강수 여부, 강설 여부, 지하철 이용자 수, 연료 가격을 지정하여 실시하였다.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	9328534947	1865706989	98.49	<.0001
Error	82	1553369257	18943528		
Corrected Total	87	10881904204			

<표2-7> ANOVA TABLE of 1st Model

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-1950.89771	26415	-0.07	0.9413	0
DdecideCnt	1	-5.28769	3.24667	-1.63	0.1072	1.18581
dummy_rain	1	3358.44777	2168.26176	1.55	0.1253	1.17038
dummy_snow	1	-5126.97769	1227.03034	-4.18	<.0001	1.22830
users	1	3.58738	0.20225	17.74	<.0001	1.27485
fuel	1	37.84247	18.07251	2.09	0.0394	1.63927

<표2-8> Parameter Estimates of 1st Model

그 결과,

회귀 직선은 $\hat{y}=-1950.89771-5.28769x_1+3358.44777x_2-5126.97769x_3+3.58738x_4+37.84247x_5$와 같이 추정되었다. (여기서 x_1은 일일 코로나 확진자 수, x_2는 강수 여부, x_3는 강설 여부, x_4는 지하철 이용자 수, x_5는 연료 가격을 나타낸다.) 즉, 눈과 비가 오지 않는 경우 (x_2=0, x_3=0)에는 추정된 회귀 직선이 $\hat{y}=-1950.89771-5.28769x_1+3.58738x_4+37.84247x_5$와 같고, 비가 오는 경우 ($x_2$=1, x_3

=0)에는 $\hat{y}= 1407.55 - 5.28769x_1 + 3.58738x_4 + 37.84247x_5$, 눈이 오는 경우 ($x_2$=0, x_3=1)에는 $\hat{y}= -7077.8754 - 5.28769x_1 + 3.58738x_4 + 37.84247x_5$와 같다고 할 수 있다. 또한 이 모형의 모형 설명력은 R_{adj}=84.85%가 나왔다. 이 회귀분석의 F=98.49 (p-value<0.0001)이므로 유의수준 5%에서 유의하다고 할 수 있다. 하지만 각각의 회귀계수를 살펴보았을 때, Intercept와 x_1, x_2에 대해서는 5% 유의수준에서 유의하지 않다는 결과가 나타났다.

위와 같이 몇몇 회귀계수가 유의하지 않은 결과가 나타났기 때문에 다중공선성을 먼저 확인해 보았다. VIF는 <표2-8>과 같이 5보다 크지 않은 값들이 나타나며, <표2-9>의 condition indices 또한 1000을 넘는 값이 존재하지 않으므로 다중공선성이 심각한 문제가 아님을 알 수 있었다. 추가로 아래의 <표2-10> correlation matrix를 통해 각 설명변수들의 상관계수가 크지 않으므로 다중공선성이 심각한 문제가 아님을 다시 한 번 확인할 수 있었다.

Collinearity Diagnostics								
			Proportion of Variation					
Number	Eigenvalue	Condition Index	Intercept	DdecideCnt	dummy_rain	dummy_snow	users	fuel
1	4.20824	1.00000	0.00001695	0.00417	0.00358	0.01227	0.00329	0.00001682
2	0.94539	2.10982	2.056241E-7	0.00028765	0.82061	0.01706	0.00049238	1.138901E-7
3	0.70968	2.43511	0.00001119	0.00091657	0.01774	0.78688	0.00305	0.00001250
4	0.10287	6.39601	0.00000129	0.44742	0.00406	0.02421	0.25781	0.00000750
5	0.03367	11.17966	0.00223	0.45507	0.05917	0.01164	0.61503	0.00216
6	0.00015411	165.24675	0.99774	0.09214	0.09484	0.14793	0.12032	0.99781

<표2-9> Collinearity Diagnostics of 1st Model

피어슨 상관 계수, N = 88 H0: Rho=0 가정하에서 Prob > |r|					
	DdecideCnt	dummy_rain	dummy_snow	users	fuel
DdecideCnt	1.00000	-0.06755 0.5318	0.20291 0.0580	-0.19644 0.0666	-0.38222 0.0002
dummy_rain	-0.06755 0.5318	1.00000	-0.01597 0.8826	-0.19242 0.0725	0.22600 0.0342
dummy_snow	0.20291 0.0580	-0.01597 0.8826	1.00000	-0.00497 0.9633	-0.38764 0.0002
users	-0.19644 0.0666	-0.19242 0.0725	-0.00497 0.9633	1.00000	0.33100 0.0016
fuel	-0.38222 0.0002	0.22600 0.0342	-0.38764 0.0002	0.33100 0.0016	1.00000

<표2-10> Correlation Matrix of 1st Model

2.2.2. 두 번째 모형

앞에서 다중공선성으로 인해 일일 코로나 확진자 수와 강수 여부에 따른 회귀계수가 0이 되는 것이 아님을 밝혀내었다. 따라서 위의 두 변수를 제거하고 다시 한 번 모델 적합을 시도해보았다.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	9231635577	3077211859	156.63	<.0001
Error	84	1650268627	19646055		
Corrected Total	87	10881904204			

<표2-11> ANOVA TABLE of 2nd Model

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-28809	24025	-1.20	0.2338	0
dummy_snow	1	-5036.84011	1235.59986	-4.08	0.0001	1.20098
users	1	3.52023	0.19529	18.03	<.0001	1.14608
fuel	1	54.95332	16.69403	3.29	0.0015	1.34872

<표2-12> Parameter Estimates of 2nd Model

두 번째 모형에 대한 적합 결과 $\hat{y}=-28809-5036.84x_1+3.52x_2+54.95x_3$의 회귀 직선을 얻을 수 있었다. (여기서 x_1은 강설 여부, x_2는 지하철 이용자 수, x_3는 연료 가격에 해당한다.) 이 때, 눈이 오지 않는 경우(x_1=0) $\hat{y}=-28809+3.52x_2+54.95x_3$와 같고, 눈이 오는 경우($x_1$=1)에는 $\hat{y}=33845.84+3.52x_2+54.95x_3$와 같아진다.

이 모형의 모형 설명력은 R_{adj}=84.29%로 두 개의 설명변수를 제거했음에도 불구하고 크게 낮아지지 않았다. 또한 이 모형의 F=156.63 (p-value<0.0001)로 유의수준 5%에서 이 모형이 유의함을 알 수 있다. 또한 Intercept를 제외한 모든 설명변수에 대한 회귀계수가 유의수준 5%에서 유의하다는 점을 알 수 있다.

2.2.3. 세 번째 모형

두 번째 모형의 결과 Intercept가 없는 모형을 사용하는 방안을 고려해볼 수 있다. 이에 따라 두 번째 모형과 같은 설명변수를 이용하는 반면 Intercept가 없는 모형을 적합해보았다. 이는 reg 프로시저의 noint 옵션을 통해 적용하였다.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	5.896742E11	1.965581E11	9953.68	<.0001
Error	85	1678518967	19747282		
Uncorrected Total	88	5.913527E11			

<표2-13> ANOVA TABLE of 3rd Model

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
dummy_snow	1	-5656.31276	1125.31360	-5.03	<.0001
users	1	3.58914	0.18712	19.18	<.0001
fuel	1	34.98356	1.16931	29.92	<.0001

<표2-14> Parameter Estimates of 3rd Model

세 번째 모형 적합 결과 $\hat{y} = -5656.31x_1 + 3.59x_2 + 34.98x_3$과 같이 회귀 직선이 추정되었다. 이는 눈이 오는 경우(x_1=0)에는 $\hat{y} = 3.59x_2 + 34.98x_3$와 같고, 눈이 오지 않는 경우($x_1$=1)에는 $\hat{y} = -5656.31 + 3.59x_2 + 34.98x_3$와 같다. <표2-13>에서 보는 것과 같이 F=9953.68(p-value<0.0001)로 추정된 회귀 직선이 유의수준 5%에서 유의함을 알 수 있다. 또한 ,이 모형의 설명력은 R_{adj}=99.71%로 매우 높게 나타났다. 회귀 직선의 절편항이 존재하지 않는 경우 R-Square 값이 높게 계산되는 점을 고려해도 매우 높은 수치이다. 마지막으로 각 회귀계수가 모두 유의수준 5%에서 유의함을 <표2-14>를 통해서 알 수 있다.

Output Statistics														
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	Cook's D	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS		
												dummy_snow	users	fuel
1	51641	58986	1198	-7346	4279.4	-1.717	0.077	-1.7368	0.0726	1.0051	-0.4860	0.0891	0.4382	-0.4820
2	63064	64177	963.6637	-1113	4338.0	-0.257	0.001	-0.2551	0.0470	1.0847	-0.0567	0.0133	0.0478	-0.0549
3	55670	61026	1109	-5356	4303.3	-1.245	0.034	-1.2487	0.0622	1.0456	-0.3217	0.0645	0.2843	-0.3171
4	82171	81105	523.3353	1065	4412.9	0.241	0.000	0.2400	0.0139	1.0486	0.0285	-0.0133	0.0034	0.0056
5	84266	81293	524.5433	2972	4412.7	0.674	0.002	0.6713	0.0139	1.0340	0.0798	-0.0373	0.0100	0.0153
6	83716	83432	543.9598	283.5128	4410.4	0.064	0.000	0.0639	0.0150	1.0517	0.0079	-0.0036	0.0021	0.0004
7	82442	84018	551.9323	-1576	4409.4	-0.357	0.001	-0.3555	0.0154	1.0476	-0.0445	0.0201	-0.0137	-0.0002
8	74949	70564	750.2764	4385	4380.0	1.001	0.010	1.0011	0.0285	1.0293	0.1715	-0.0537	-0.1224	0.1538
9	84728	83814	546.9841	913.9233	4410.0	0.207	0.000	0.2061	0.0152	1.0505	0.0256	-0.0117	0.0069	0.0011
10	85696	85515	576.5786	180.8280	4406.2	0.041	0.000	0.0408	0.0168	1.0538	0.0053	-0.0023	0.0022	-0.0005
11	84595	85752	580.4795	-1157	4405.7	-0.263	0.000	-0.2611	0.0171	1.0516	-0.0344	0.0150	-0.0144	0.0039
12	77812	72851	681.3248	4961	4391.3	1.130	0.010	1.1317	0.0235	1.0140	0.1756	-0.0613	-0.1111	0.1481
13	65666	66069	945.4000	-402.6496	4342.1	-0.093	0.000	-0.0922	0.0453	1.0849	-0.0201	0.0049	0.0167	-0.0193
14	86135	85913	583.1652	221.5133	4405.4	0.050	0.000	0.0500	0.0172	1.0542	0.0066	-0.0029	0.0028	-0.0008
15	83957	85238	568.9442	-1281	4407.2	-0.291	0.000	-0.2891	0.0164	1.0503	-0.0373	0.0165	-0.0139	0.0024
16	87866	86741	602.5681	1125	4402.7	0.256	0.000	0.2541	0.0184	1.0531	0.0348	-0.0146	0.0168	-0.0064
17	84750	86306	590.5108	-1556	4404.4	-0.353	0.001	-0.3513	0.0177	1.0501	-0.0471	0.0202	-0.0210	0.0068
18	86582	86117	586.1507	464.7109	4405.0	0.105	0.000	0.1049	0.0174	1.0541	0.0140	-0.0060	0.0060	-0.0018
19	78727	74437	640.2010	4290	4397.4	0.976	0.007	0.9753	0.0208	1.0230	0.1420	-0.0533	-0.0800	0.1127
20	66604	67157	910.0842	-552.6348	4349.6	-0.127	0.000	-0.1263	0.0419	1.0809	-0.0264	0.0067	0.0215	-0.0252
21	86134	87498	617.6494	-1364	4400.7	-0.310	0.001	-0.3084	0.0193	1.0530	-0.0433	0.0179	-0.0223	0.0096
22	88146	88060	631.1035	86.2627	4398.7	0.020	0.000	0.0195	0.0202	1.0575	0.0028	-0.0011	0.0015	-0.0007
23	86112	86603	591.2833	-491.9343	4404.3	-0.112	0.000	-0.1110	0.0177	1.0544	-0.0149	0.0064	-0.0066	0.0021
24	63745	65093	1015	-1348	4326.2	-0.312	0.002	-0.3100	0.0522	1.0895	-0.0728	0.0165	0.0620	-0.0708
25	60240	60953	1206	-712.8904	4277.0	-0.167	0.001	-0.1657	0.0737	1.1174	-0.0467	0.0087	0.0419	-0.0463
26	62483	67265	923.0347	-4783	4346.9	-1.100	0.018	-1.1016	0.0431	1.0373	-0.2339	0.0590	0.1912	-0.2238
27	62090	66915	938.8067	-4825	4343.5	-1.111	0.019	-1.1124	0.0446	1.0380	-0.2404	0.0596	0.1982	-0.2308
28	84833	86457	584.5006	-1624	4405.2	-0.369	0.001	-0.3667	0.0173	1.0493	-0.0487	0.0212	-0.0202	0.0053
29	89877	90006	671.1802	-128.7082	4392.8	-0.029	0.000	-0.0291	0.0228	1.0603	-0.0045	0.0017	-0.0027	0.0014
30	81421	76994	602.4448	4427	4402.8	1.006	0.006	1.0057	0.0184	1.0183	0.1376	-0.0559	-0.0632	0.0983
31	70246	69646	851.2536	599.5870	4361.5	0.137	0.000	0.1367	0.0367	1.0749	0.0267	-0.0074	-0.0207	0.0249
32	88216	89040	637.8548	-824.2045	4397.8	-0.187	0.000	-0.1863	0.0206	1.0566	-0.0270	0.0110	-0.0147	0.0068
33	90666	89182	640.0237	1484	4397.5	0.337	0.001	0.3357	0.0207	1.0538	0.0489	-0.0198	0.0267	-0.0125
34	90641	89659	651.9344	981.8911	4395.7	0.223	0.000	0.2221	0.0215	1.0571	0.0329	-0.0131	0.0187	-0.0093
35	90950	89582	648.0183	1367	4396.3	0.311	0.001	0.3093	0.0213	1.0550	0.0456	-0.0183	0.0255	-0.0124
36	90995	91101	693.2744	-105.6159	4389.4	-0.024	0.000	-0.0239	0.0243	1.0620	-0.0038	0.0014	-0.0024	0.0013
37	84524	78564	581.9819	5960	4405.5	1.353	0.011	1.3595	0.0172	0.9876	0.1796	-0.0764	-0.0682	0.1168
38	73646	71488	800.0294	2158	4371.2	0.494	0.003	0.4914	0.0324	1.0617	0.0899	-0.0270	-0.0665	0.0821
39	89864	90603	666.5535	-739.1106	4393.5	-0.168	0.000	-0.1673	0.0225	1.0589	-0.0254	0.0100	-0.0149	0.0077
40	90512	90473	660.6107	38.5959	4394.4	0.009	0.000	0.008731	0.0221	1.0596	0.0013	-0.0005	0.0008	-0.0004
41	91310	92837	734.7092	-1527	4382.6	-0.348	0.001	-0.3467	0.0273	1.0607	-0.0581	0.0209	-0.0393	0.0239
42	82423	79066	589.3999	3357	4404.5	0.762	0.003	0.7602	0.0176	1.0332	0.1017	-0.0431	-0.0397	0.0670

<표2-15> Residual Diagnostics of 3rd Model (계속)

43	70952	71386	835.5189	-434.8267	4364.5	-0.100	0.000	-0.0990	0.0354	1.0737	-0.0190	0.0055	0.0144	-0.0175
44	89082	90751	656.4271	-1669	4395.0	-0.380	0.001	-0.3779	0.0218	1.0539	-0.0564	0.0226	-0.0315	0.0153
45	91328	91525	674.2740	-197.1211	4392.3	-0.045	0.000	-0.0446	0.0230	1.0605	-0.0068	0.0027	-0.0040	0.0021
46	91108	92108	688.3024	-999.5618	4390.2	-0.228	0.000	-0.2264	0.0240	1.0597	-0.0355	0.0137	-0.0216	0.0117
47	90696	91767	674.2768	-1071	4392.3	-0.244	0.000	-0.2425	0.0230	1.0583	-0.0372	0.0146	-0.0217	0.0112
48	90908	92942	708.3473	-2035	4387.0	-0.464	0.002	-0.4617	0.0254	1.0551	-0.0745	0.0281	-0.0472	0.0267
49	83079	80563	586.5961	2516	4404.9	0.571	0.002	0.5689	0.0174	1.0424	0.0758	-0.0327	-0.0265	0.0474
50	71258	72428	839.9425	-1170	4363.7	-0.268	0.001	-0.2665	0.0357	1.0718	-0.0513	0.0149	0.0388	-0.0473
51	87574	91457	652.8946	-3884	4395.6	-0.884	0.006	-0.8824	0.0216	1.0301	-0.1311	0.0534	-0.0703	0.0322
52	92882	92193	669.8013	688.9476	4393.0	0.157	0.000	0.1559	0.0227	1.0593	0.0238	-0.0095	0.0135	-0.0066
53	94348	92869	686.3710	1479	4390.5	0.337	0.001	0.3352	0.0239	1.0572	0.0524	-0.0205	0.0310	-0.0162
54	95334	92880	683.6985	2453	4390.9	0.559	0.003	0.5564	0.0237	1.0496	0.0866	-0.0340	0.0508	-0.0263
55	95939	94926	747.6493	1012	4380.4	0.231	0.001	0.2298	0.0283	1.0643	0.0392	-0.0142	0.0263	-0.0159
56	73960	73057	851.9343	903.2735	4361.4	0.207	0.001	0.2059	0.0368	1.0740	0.0402	-0.0116	-0.0305	0.0372
57	92646	92121	654.2003	525.2433	4395.4	0.119	0.000	0.1188	0.0217	1.0586	0.0177	-0.0072	0.0093	-0.0041
58	95671	93026	678.3808	2644	4391.7	0.602	0.003	0.5998	0.0233	1.0473	0.0927	-0.0368	0.0530	-0.0265
59	95897	93392	688.1390	2504	4390.2	0.570	0.003	0.5681	0.0240	1.0495	0.0890	-0.0349	0.0523	-0.0271
60	91739	93671	696.2147	-1933	4388.9	-0.440	0.002	-0.4383	0.0245	1.0550	-0.0695	0.0270	-0.0417	0.0222
61	95415	95137	743.8958	278.2518	4381.1	0.064	0.000	0.0631	0.0280	1.0659	0.0107	-0.0039	0.0071	-0.0042
62	70231	92157	650.2818	-21926	4396.0	-4.988	0.181	-5.8956	0.0214	0.3747	-0.8721	0.3603	-0.4490	0.1924
63	94624	93291	681.9838	1332	4391.1	0.303	0.001	0.3018	0.0236	1.0577	0.0469	-0.0186	0.0270	-0.0136
64	95493	93836	698.3922	1657	4388.6	0.378	0.001	0.3756	0.0247	1.0570	0.0598	-0.0232	0.0360	-0.0192
65	85281	84767	561.4003	513.5358	4408.2	0.116	0.000	0.1158	0.0160	1.0524	0.0147	-0.0066	0.0051	-0.0005
66	82842	78798	653.1246	4044	4395.5	0.920	0.006	0.9191	0.0216	1.0277	0.1366	-0.0529	-0.0719	0.1046
67	87400	79176	652.8352	8224	4395.6	1.871	0.026	1.8994	0.0216	0.9335	0.2821	-0.1097	-0.1470	0.2149
68	72331	71278	936.4829	1053	4344.0	0.242	0.001	0.2410	0.0444	1.0820	0.0520	-0.0136	-0.0418	0.0493
69	73455	76789	998.5446	-3335	4330.1	-0.770	0.011	-0.7683	0.0505	1.0686	-0.1772	-0.1559	-0.0175	0.0170
70	58003	78345	1009	-20342	4327.7	-4.700	0.400	-5.4316	0.0516	0.4428	-1.2663	-1.1002	-0.2196	0.2110
71	69893	75703	994.2364	-5811	4331.1	-1.342	0.032	-1.3481	0.0501	1.0229	-0.3095	-0.2737	-0.0104	0.0104
72	64738	62160	1203	2578	4278.0	0.603	0.010	0.6002	0.0732	1.1037	0.1687	0.1255	-0.0950	0.0900
73	54803	56373	1396	-1571	4218.8	-0.372	0.005	-0.3704	0.0987	1.1440	-0.1226	-0.0791	0.0861	-0.0816
74	79675	76253	995.0665	3422	4330.9	0.790	0.011	0.7884	0.0501	1.0670	0.1811	0.1599	0.0096	-0.0094
75	74252	78850	1011	-4598	4327.3	-1.063	0.021	-1.0634	0.0517	1.0497	-0.2483	-0.2151	-0.0453	0.0433
76	76157	78380	1006	-2223	4328.5	-0.514	0.005	-0.5114	0.0512	1.0820	-0.1188	-0.1035	-0.0185	0.0177
77	58849	58230	1347	619.3825	4234.7	0.146	0.001	0.1454	0.0919	1.1401	0.0463	0.0308	-0.0312	0.0296
78	72823	79395	1012	-6572	4327.0	-1.519	0.042	-1.5307	0.0519	1.0062	-0.3580	-0.3092	-0.0678	0.0646
79	82222	80120	1017	2101	4325.9	0.486	0.004	0.4835	0.0524	1.0843	0.1137	0.0975	0.0241	-0.0229
80	87332	79804	1013	7527	4326.7	1.740	0.055	1.7611	0.0520	0.9803	0.4124	0.3553	0.0807	-0.0766
81	83520	81968	1040	1552	4320.3	0.359	0.002	0.3574	0.0548	1.0912	0.0861	0.0720	0.0255	-0.0242
82	81763	81986	1040	-223.5911	4320.3	-0.052	0.000	-0.0514	0.0548	1.0961	-0.0124	-0.0104	-0.0037	0.0035
83	87356	82784	1053	4572	4317.2	1.059	0.022	1.0597	0.0561	1.0549	0.2585	0.2134	0.0855	-0.0812
84	84530	82103	1032	2426	4322.3	0.561	0.006	0.5590	0.0539	1.0831	0.1335	0.1124	0.0361	-0.0340
85	88052	81664	1025	6388	4324.0	1.477	0.041	1.4878	0.0532	1.0122	0.3527	0.2991	0.0864	-0.0815
86	88688	82061	1029	6626	4323.0	1.533	0.044	1.5453	0.0536	1.0065	0.3679	0.3105	0.0956	-0.0901
87	87700	84295	1049	3405	4318.2	0.788	0.012	0.7867	0.0557	1.0734	0.1911	0.1575	0.0612	-0.0574
88	65206	61746	1298	3459	4249.9	0.814	0.021	0.8123	0.0854	1.1066	0.2482	0.1701	-0.1597	0.1524

<표2-15> Residual Diagnostics of 3rd Model

한남대교의 교통량에 대한 충분히 높은 모형 설명력을 보이는 모형을 찾았으므로 influential point 분석을 실시하였다. <표2-15>와 같이 각 관측치의 특성을 파악할 수 있는 잔차를 비롯한 다양한 값들을 구할 수 있다. 여기서 Rstudent Residual의 절대값이 2.5보다 큰 값을 이상치로 판단하였다. 62번째, 70번째 관측치가 이에 해당한다.

2.2.4. 네 번째 모형

이상치로 판단된 62번째, 70번째 관측치를 데이터셋에서 제거하고 세 번째 모형과 같은 모형으로 적합을 시도해보았다.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	5.823078E11	1.941026E11	21530.5	<.0001
Error	83	748266518	9015259		
Uncorrected Total	86	5.83056E11			

<표2-16> ANOVA TABLE of 4th Model

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
dummy_snow	1	-4925.56237	777.14525	-6.34	<.0001
users	1	3.69608	0.12690	29.13	<.0001
fuel	1	34.57857	0.79109	43.71	<.0001

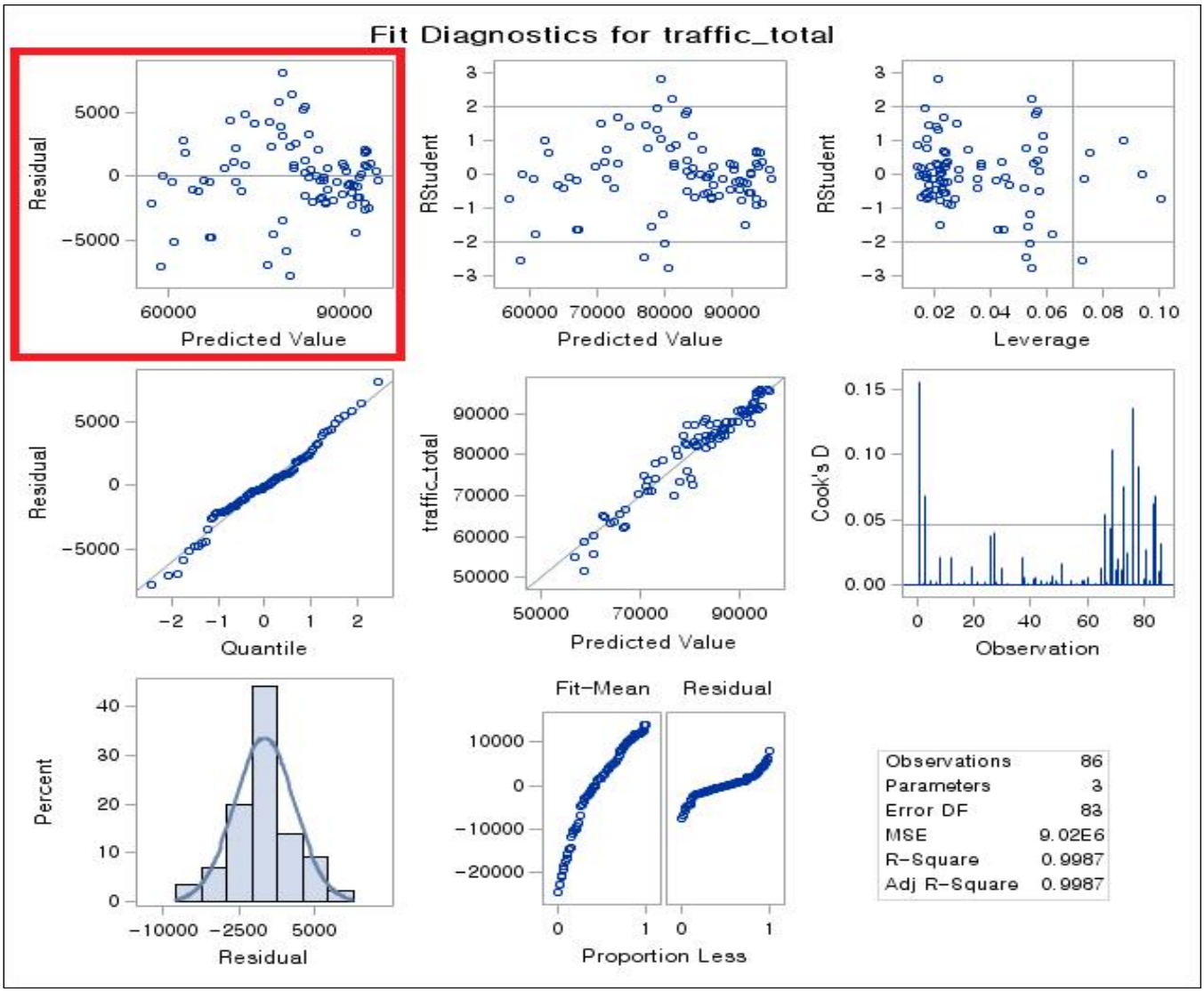
<표2-17> Parameter Estimates of 4th Model

적합 결과 $\hat{y}=-4925.56x_1+3.7x_2+34.58x_3$과 같은 회귀 직선을 얻을 수 있다. 이는 눈이 오는 경우인 경우(x_1=0) $\hat{y}=3.7x_2+34.58x_3$, 눈이 오지 않는 경우(x_1=1), $\hat{y}=-4925.56+3.7x_2+34.58x_3$과 같다. 이 모형에 대한 F=21530.5(p-value<0.0001)로 모형 적합이 유의수준 5%에서 유의하며, 각 회

귀계수들 또한 같이 유의수준에서 유의함을 알 수 있다. 이 모형의 모형 설명력은 R_{adj}=99.87%로 이상치를 제거하지 않은 세 번째 모형보다 소폭 상승했으며 매우 높은 설명력을 보여준다. 100%에 가까운 모형 설명력을 보이는 모형을 찾아냈고, 이상치 또한 제거하는 과정을 거쳤으므로 이 모형의 타당성 점검, 즉 모형 가정이 만족하는지 확인하는 단계로 넘어간다.

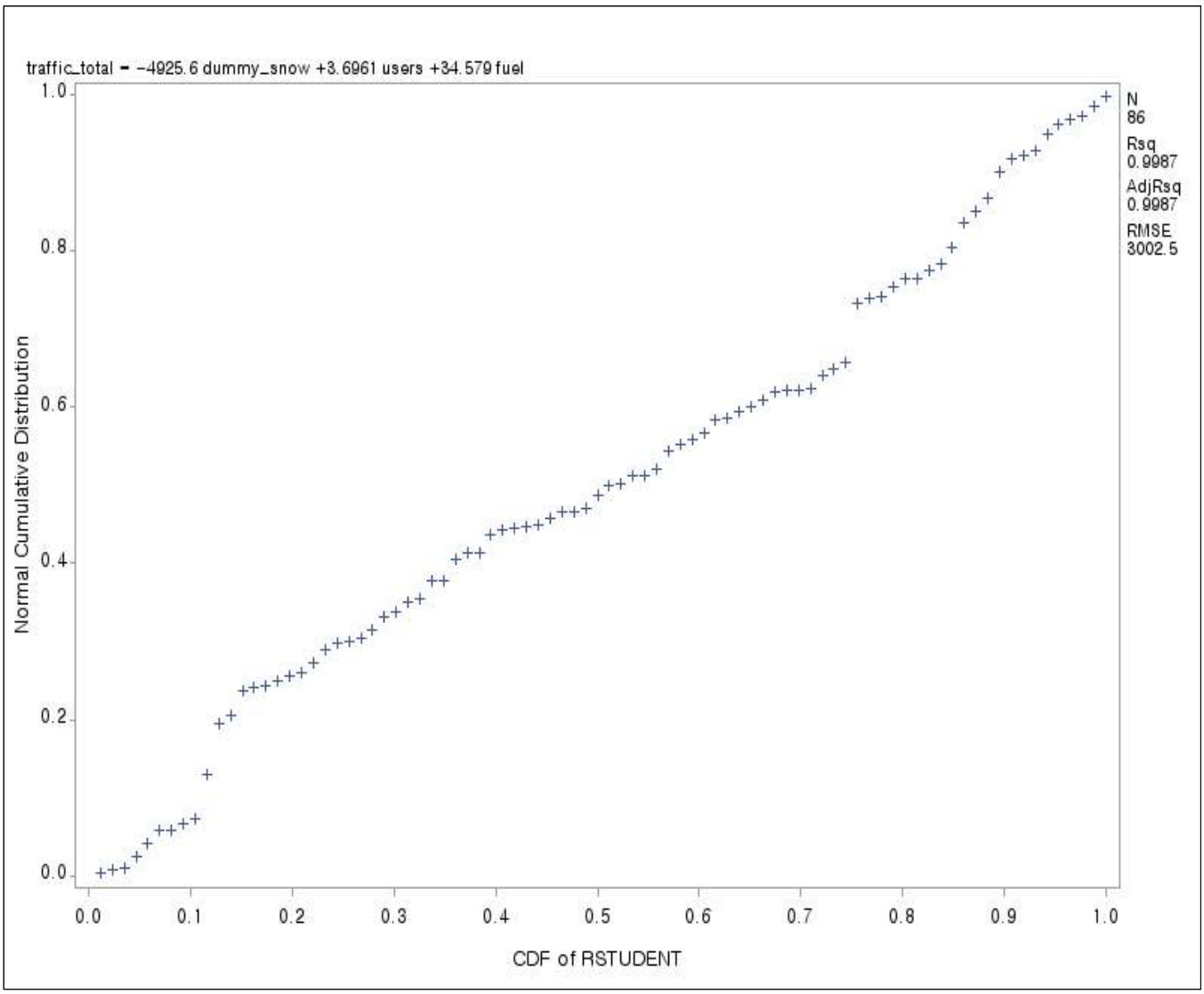
2.3. 모형 타당성 점검

가장 높은 설명력을 보이는 마지막 모형부터 모형 타당성을 점검한다. 만약 점검한 모형이 타당하지 않다는 결론이 도출되는 경우에는 그 전 모형을 점검하는 방식으로 진행한다. 모형 타당성은 Residual Plot을 통해 등분산성 가정이 만족되는지, Normal Probability Plot을 통해 정규성 가정이 만족되는지를 확인한다.



<그래프2-6> fit diagnostics

위 <그래프2-6>에서 Residual Plot을 살펴보았을 때, 특이점은 없는 것으로 보인다. 따라서 마지막으로 고려한 모형, 즉 intercept가 존재하지 않고, 강설 여부, 지하철 사용자수, 연료 가격을 고려했으며 이상치를 제거한 모형에서 등분산성 가정에 결함이 없다고 판단할 수 있다.



<그래프2-7> normal probability plot

<그래프2-7>는 nmormal probability plot을 확대해서 그려본 결과이다. 위 그래프를 봤을 때 대부분의 점이 직선 근처에 위치하므로 정규성 가정에 크게 위배되지 않는 것으로 판단할 수 있다.

Residual Plot에서 특별한 패턴이 발생하지 않았다는 점을 바탕으로 등분산성 가정 충족되었으며, Normal probability plot에서 점들이 직선 근처에 분포한다는 점을 통해 정규성 가정이 위배되지 않음을 확인했다. 따라서 고려한 모형의 적합도가 적절하다고 판단할 수 있다.

3. 결론

3.1. 분석 결과 요약

한남대교 교통량에 영향을 끼칠 수 있는 요인으로 총 5가지를 선정하여 탐색적 자료 분석부터 회귀 분석까지 실시하였다. 5가지 요인으로는 일일 코로나 확진자 수, 강수 여부, 강설 여부, 지하철 이용자 수, 연료 가격이 포함되었다.

본격적인 회귀 분석에 들어가기 앞서 각 자료들의 특성과 변수들의 관계를 알아보는 탐색적 자료 분석 단계를 거쳤다. 수치형 변수와 더미 변수를 구분하여 각자 다른 방식으로 실시하였는데, 수치형 변수는 각 변수들의 요약 통계량을 먼저 계산해보고, 반응변수와 각 설명변수들의 상관관계를 파악하여 시각화해보았다. 이를 통해 어떠한 설명변수가 반응변수에 어떤 영향을 끼치는 예상해볼 수 있었다. 한편, 더미 변수에 대해서는 먼저 더미 변수의 값에 따른 반응변수의 요약통계량을 계산했다. 이 과정에서 강수, 강설 여부가 교통량에 영향을 끼칠 가능성이 존재한다는 점을 발견하였다. 또한 이 요약통계량을 boxplot을 이용해 시각화하면서 시각적으로 더 받아들이기 쉽게 만들었다.

탐색적 자료 분석 단계를 거쳐 본격적인 회귀 분석 단계로 들어가보았다. 우선적으로, 처음 고려한 모든 설명 변수를 포함한 다항 회귀 모형에 대해 회귀 분석을 실시했다. 그 결과 일일 코로나 확진자 수와 강수 여부에 대한 회귀계수가 통계적으로 유의하지 않다는 결론이 나왔다. 이에 따라 이 둘을 제외한 새로운 모형에 적합을 시도했다. 설명변수들에 대한 회귀계수는 통계적으로 유의하지만 절편항이 유의하지 않다는 결과가 두 번째 모형에서 나타났다. 이를 바탕으로 두 번째 모형과 같이 강설 여부, 지하철 이용자 수, 연료 가격을 설명변수로 사용하고 절편항을 제거한 세 번째 모형을 적합해보았다. 그 결과 이 모형의 매우 높은 설명력과 회귀분석, 각 회귀계수들이 유의하다는 점을 알 수 있었다.

별 다른 변환 등을 거치지 않고 충분히 높은 설명력을 보여주는 모형을 찾아냈으므로 이 모형에 대한 influence point를 진단하는 과정을 진행하였다. 여기서 2개의 이상치를 탐색해내었고 이를 제거하고 다시 한 번 모형을 적합하였다. 이상치를 제거함으로서 모형 설명력을 조금 더 높여주었다.

마지막으로 이 모형의 가정이 충분히 충족되는지 확인하는 모형 타당성 점검을 실시했다. 크게 두 가지 부분에 대해 이를 진행했는데, 등분산성 가정과 정규성 가정이 이에 해당한다. Residual Plot에서 어떠한 패턴이 나타나지 않는 점을 근거로 등분산성 가정이 충족된다는 판단을 내렸으며, Normal probability plot에서 점들이 y=x에 해당하는 직선 근처에 분포해있다는 점을 바탕으로 정규성 가정이 만족된다고 결정하였다.

3.2. 분석의 장점 및 한계점

실시한 분석의 장점을 한 가지 뽑자면 단순한 모형을 통해 한남대교의 교통량을 분석했다는 점이다. 다소 복잡할 수 있다고 느낄 수 있는 주제이지만 단순한 모형을 사용함으로서 해석이 직관적으로 이해하는데 어려움이 없다. 즉, 단순한 모형을 통한 직관적 이해가 위 분석의 가장 큰 장점이라고 할 수 있다.

이 분석의 한계점으로는 크게 3가지를 뽑을 수 있다. 우선, 과연 교통량 분석에 다중 회귀 모형이 적절한가에 대한 점이다. 단순한 모형으로 해석이 용이하다는 장점이 존재하지만 교통량은 Count data임을 고려했을 때 포아송 회귀 모형 혹은 음이항 회귀 모형이 더 적절할 수 있을 것이다.

두 번째로는 31개의 한강 대교 중 한남대교만을 대상으로 했다는 점이다. 물론 한남대교만을 고려해 분석하는 것도 분명 가치 있지만 분석이 범용적으로 차용되기 위해서는 모든 31개 한강대교, 혹은 31개 중 많은 샘플을 바탕으로 한 분석이 더 필요하다.

마지막 한계점으로는 더 많은 요인을 고려하지 못했다는 점이다. 다중 회귀 모형에서 많은 설명변수를 사용하는 것이 무조건적으로 좋은 것은 아니지만 조금 더 다양한 요인을 고려하고 그 중에서 추려나가는 과정을 거쳤다면 더 좋은 요인을 찾을 수 있을 것 같다는 아쉬움이 남는다.

3.3. 추가 연구사항 제안

추가 연구사항으로 고려해볼 수 있는 사항으로 크게 세 가지를 제안할 수 있다. 먼저, 한남대교만을 바탕으로 하는 것에서 31개의 한강 대교에 대한 분석으로 확장하는 것이다. 앞서 진행한 연구의 한계점으로 지적한 점을 추가적인 연구를 통해 확장한다면 더 범용적인 분석 결과를 도출해낼 수 있을 것으로 기대된다.

두 번째로는 새로운 모형을 고려하는 것이다. 이 또한 연구의 한계점으로 지적했던 점으로 포아송 회귀 모형 혹은 음이항 회귀 모형을 사용해 회귀 분석을 실시한다면 연구에 사용하는 자료의 특성 상 모형의 적합도가 더 좋을 것이라고 예상된다.

마지막으로는 추가적인 설명변수 후보의 추가 및 데이터 양 증가이다. 위 연구에서는 5가지 요인 밖에 고려를 하지 못했다. 특히, 기상 정보 중 안개에 대한 데이터가 너무 적어서 이용하지 못했던 점에 대해 아쉬움이 남는다. 생각해보지 못했던 새로운 요인들을 설명변수로 사용해서 다양한 모형을 적합하는 기회를 갖는다면 좋을 것 같다. 또한 2021년 1월부터 3월까지의 데이터만 이용하는 것이 아니라 2018년부터 현재까지의 데이터를 사용하는 것과 같은 방식으로 더 많은 데이터를 사용한다면 더 정확한 추론이 이루어질 수 있을 것이다.

4. 참고문헌

[1] 김성수. 『대도시의 대중교통수요 영향요인 분석』 서울대학교 대학원 : 환경대학원 환경계획학과, 2019. 2.

[2] 월별 서울시 교통량 조사자료와 2020년 서울특별시 교통량 조사자료

: TOPIS - 교통량 정보 : https://topis.seoul.go.kr/refRoom/openRefRoom_2.do

[3] 코로나 확진자 수 -> KDX 한국데이터거래소 https://kdx.kr/main

[4] 기상 자료

: 기상자료개방포털 : https://data.kma.go.kr/data/grnd/selectAsosRltmList.do?pgmNo=36

[5] 지하철 역별 이용자 수

: 서울 열린 데이터 광장 : http://data.seoul.go.kr/dataList/OA-12914/S/1/datasetView.do

[6] 서울시 자동차 등록현황(연료별) 통계

: 서울 열린 데이터 광장 : https://data.seoul.go.kr/dataList/10860/S/2/datasetView.do

[7] 주유소_평균판매가격_제품별.csv

: Opinet : https://www.opinet.co.kr/user/dopospdrg/dopOsPdrgSelect.do#

[8]「비 오면 교통량 3.1% 감소...봄철 주말에 가장 민감」, 한국도로공사 보도자료