**회귀분석I 기말과제**

<서울특별시 상권의 특징에 따른 매출추정>

담당교수 : 김규성 교수님

제출일자 : 2022.06.17

**2019580027 정우진**

목차

**1. 서론3**

1.1. 연구 목적2

1.2. 문헌 연구 3

1.3. 데이터 설명 3

1.4 분석 방법3

1.5. 결과 활용 및 기대 효과3

**2. 본론4**

2.1 분석 방법 소개5

2.2 데이터 분석 및 결과 설명6

2.3 분석의 타당성 설명6

**3. 결론4**

3.1 분석 결과 요약5

3.2 분석의 장점 및 한계점 설명18

3.3. 추가 연구사항 제안19

참고문헌**20**

부록(SAS CODE)**20**

**1. 서론**

**1.1 연구 목적**

상점이 위치한 상권이 그 상점의 매출에 큰 영향을 미친다는 것은 당연하게 받아들여진다. 소비자들에게 동일한 서비스나 재화를 제공하는 상점이라도, 소위 말해 장사가 잘되는 상권, 가령 홍대나 강남 같은 곳에 있는 가게가 더 높은 매출을 기록할 것이다. 그렇다면, 상점들의 매출을 높이는 상권의 특징은 무엇이며, 특징과 매출간에 어떠한 관계가 존재하는가? 이 연구에서는 ‘서울 열린데이터 광장’에서 제공하는 데이터를 이용하여, 위 질문에 답하고자 한다.

**1.2. 데이터 설명**

본 연구에 사용된 데이터는 ‘서울 열린데이터 광장’의 ‘골목상권분석정보’에서 제공되는 자료 중, 상권에 따른 특징의 수준을 정리한 6개의 데이터셋 각각에서 대표적인 변수들을 가져와 상권매출에 대한 새로운 데이터셋으로 만든 것이다. 데이터셋에 포함된 변수들은 다음과 같다. 여기서 상권배후지란, 상권 경계 200m 이내 구역을 의미한다.

|  |  |
| --- | --- |
| 변수명 | 변수 설명 |
| VAR1 | 상권 명. 상권의 위치를 나타내며, 도로명으로 표기되어 있다. |
| income | 상권배후지 내 거주 인구의 월 평균 소득금액이다. 2021년 1~4분기 값의 평균이다. |
| Pay | 상권배후지 내 거주 인구의 분기 별 총 지출 금액이다. |
| subway | 상권배후지 내 지하철 역의 개수를 의미한다. |
| busstop | 상권배후지 내 버스정류장 개수를 의미한다. |
| school | 상권배후지 내 학교 수(초, 중, 고) |
| Apart\_count | 상권배후지 내 아파트 세대 수. |
| Apart\_price | 상권배후지 내 아파트 평균 가격 |
| Working\_pop | 상권배후지 내 직장인구. |
| Var10 | 상권배후지 총 생활인구 수. |
| Y | 반응변수. 상권배후지 별 상점평균 매출. |

**1.3. 분석 방법**

이 보고서에서는 다중선형회귀적합을 통하여 데이터를 분석한다.

데이터 셋

데이터 표준화

다중 회귀모형 선택

모형 타당성 검증

변환 또는 가중

회귀진단

모델링 완료

**<그림 1 : 데이터 분석 프로세스>**

우선 다중선형회귀를 이용하여 예측변수와 반응변수 간 관계를 규명한다. 이 때 변수 선택과정에서, Backward Selection 알고리즘을 활용한다. 이 때 다중공선성도 같이 확인 해준다.

적합 모형이 도출되었으면, 모형타당성 검증을 시행한다. 만약 모형타당성이 좋지 않은 것으로 판단된다면, 데이터 변형 혹은 가중절차를 시행 한 후 다시 첫 단계로 돌아간다. 모형 타당성까지 만족하는 모델이 도출 되었다면, 각 관측치의 영향력을 확인하는 회귀진단을 한 후, 모델을 확정한다.

**1.4. 결과 활용 및 기대 효과**

이 연구는 상권의 매출에 영향을 미치는 상권의 특징을 분석하는 것이 목적이다.

따라서, 이 연구의 결과는 새로운 사업장의 위치를 고민하는 상인들의 의사결정에 도움이 될 수 있다.

**2. 본론**

**2.1. 분석 방법 소개**

앞에서 간단히 설명한 바와 같이, 분석의 첫 단계에서는 Backward Selection을 이용한 다중회귀적합을 실시한다. 여기서 Backward selection이란, 모든 예측변수를 포함한 상태에서 다중선형회귀적합을 실시하여, 이 중 가장 영향력이 적다고 판단되는 변수를 하나씩 제외해 가며 선형회귀적합을 실시하는 방법이다. 이 때 영향력 판단 기준은 Adjusted – R squared, AIC, Cp 통계량을 참고하여 결정한다.

모형 타당성 검증에서는, 주로 잔차분석을 통하여 선형회귀 기본 가정들을 검증한다. 이 때, 1) 예측변수와 반응변수 간의 선형성, 오차 epsilon의 2) 등분산성, 3) 정규성, 4) 평균 = 0 을 검증한다. (시계열 데이터가 아니므로 오차의 독립성에 대한 검증은 생략한다.) 이 때 모형이 타당하지 않은 것으로 판정되었다면, 데이터 변환이나 가중 작업을 고친 후 다시 첫 단계로 돌아간다.

마지막 단계에서는 회귀진단을 실시한다. 여기서는 분석 결과의 이상치나 영향력 관측치를 관찰한다. 관찰을 통해 데이터의 수정이 필요하다고 생각될 때에는 데이터 수정 후 다시 첫 단계로 돌아간다.

모든 단계가 완료되었다면, 분석결과를 해석하고 프로세스를 종료한다.

**2.2. 데이터 분석 및 결과 설명**

1, 데이터 표준화

먼저, 데이터의 모든 변수를 표준화해주는 작업을 거쳤다. 변수들을 표준화함으로서 변수 간 불필요한 Collinearity 를 없애고, 회귀계수의 크기를 조정하며, 잔차분석 시 평균점을 잘 파악할 수 있는 등의 이점이 있다. 표준화 절차는 SAS의 standard 프로시저를 이용하였다.

/\*데이터 표준화\*/

**proc** **standard** data = market out = stdmarket mean = **0** std= **1**;

**run**;

2, 다중회귀모형 선택

다음으로, 다중회귀모형 선택 작업을 시작하였다. 먼저 Full model(모든 설명변수 포함)을 적합시켜 변수 들 사이 다중공선성 존재 여부를 확인하고자 하였다.

**proc** **reg** data = stdmarket;

model y = income pay subway busstop school apart\_count apart\_price var10 working\_pop / vif collin;

**run**; **quit**;

Variance Inflation Factor(VIF)와, eigenvalue condition number를 고려하면, 고려되는 총 10개 변수 사이엔 심각한 다중공선성이 발견되지 않았다. 따라서, 다중공선성은 앞으로 고려하지 않는다.

위 10개 변수를 사용한 모델 중 최적의 모델을 선정하기 위해, Proc Reg를 사용하여 Full model에 대하여 Backward Selection을 시행하여 변수를 선정하였다.

테이블이(가) 표시된 사진

자동 생성된 설명테이블이(가) 표시된 사진

자동 생성된 설명<그림 1 : Full model에서 다중공선성을 탐색한 결과 : 일반적으로 Vif가 10 또는 5 이상, condition number 가 100이상이면 다중공선성이 있는 것으로 본다. 본 모델의 경우는 해당사항이 없음을 알 수 있다>

<Backward Selection 코드>

/\*backward selection - 프로그램에서는 모든 계수 p-value가 0.1보다 작으면 멈춘다.(Default)\*/

**proc** **reg** data = stdmarket;

model y = income pay subway busstop school apart\_count apart\_price var10 working\_pop / selection = backward;

**run**; **quit**;

Proc reg에서는 모든 회귀계수 검정 T 통계량이 0.1보다 작으면 elimination을 멈춘다. 본인은 C(p)가 가장 작은,

즉 apart\_count까지만 제거된 모델을 선택하였다.

테이블이(가) 표시된 사진

자동 생성된 설명

<그림 2: Backward Selection 결과 요약 값>

(이 때, C(p)는 (SSres(p) / sigma\_hat) - n +2p로 표현되며, 그 값이 작을수록 모델의 적합이 좋다고 본다.) 또한, 결정한 모델에서의 busstop의 회귀계수 검정통계량의 p-value가 0.13 가량으로 일반적으로 귀무가설 H0 : B = 0을 기각하기엔 불충분하나 어느 정도의 유의성은 있다고 여겨져 4변수를 삭제한 6개의 반응변수를 가진 모델을 최적모델으로 결정하였다. 최적모델에 대한 요약통계량은 다음과 같다.

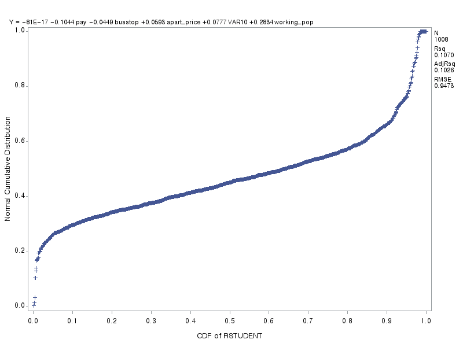
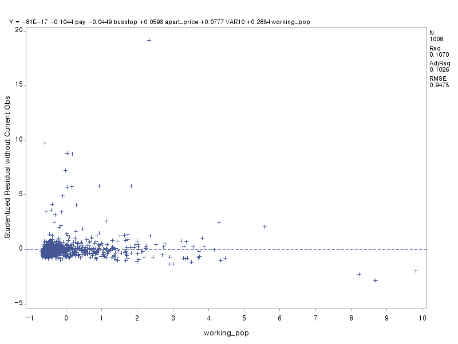
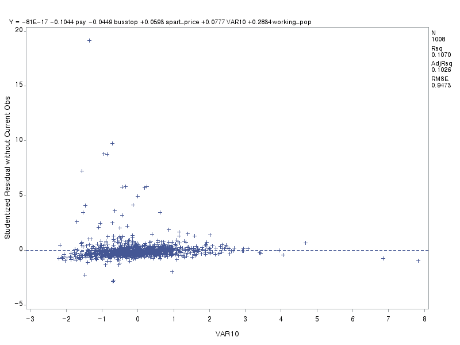
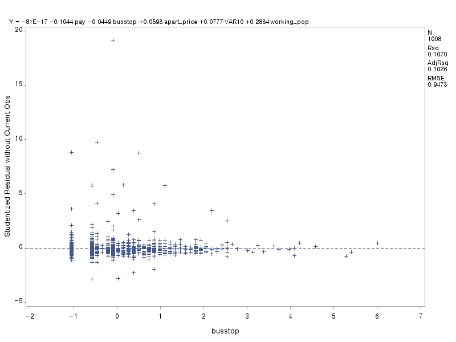
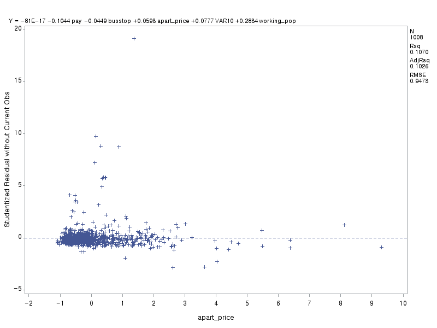
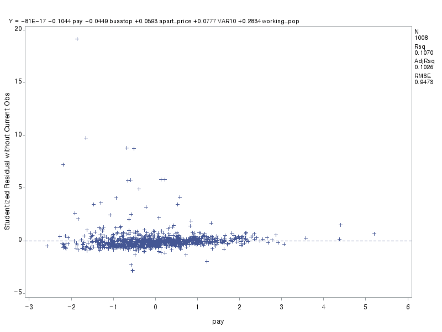
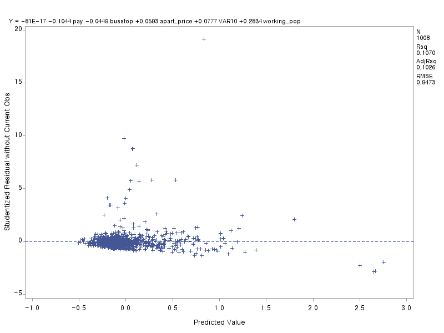
테이블이(가) 표시된 사진

자동 생성된 설명<그림 3 : 결정된 최적모델의 요약통계량>

위의 그림 3을 보면, 이 모델의 경우 ANOVA F 통계량이 매우 작으므로 H0 : B0…B6 = 0을 기각한다. 회귀계수 검정의 경우에도 busstop을 제외한 나머지 모든 변수의 계수값이 유의수준 0.1에서 귀무가설 H0 : B = 0을 기각하며, busstop의 경우에도 어느정도의 유의성은 가지는 것으로 판단된다. 이 모델의 경우, R squared 통계량이 매우 낮은 것이 눈에 띄는데, 이는 y값을 가지고 있는 설명계수들 만으로 충분히 설명하지 못함을 의미한다. (변수선택의 문제는 아닌게, full model의 r squared 값 역시 이와 비슷하다.) 다만, 서울시에서 제공한 데이터들의 핵심변수들을 대부분 분석과정에서 이용하고 있고, R-squared가 0.1이하로 낮을 지라도 크게 문제가 되지 않을 수 있다는 통계학자들[[1]](#footnote-1)의 말을 인용하여, 이를 큰 문제로 생각하지 않고 지나가겠다.

3. 모형 타당성 검증

다음으로 , 구해진 모형의 모형 타당성 검증을 수행하겠다.

<그림 4 : 회귀모델의 각 변수에 대한 잔차플롯들과 normal probability plot>

위의 그림 4를 참고하여 잔차분석을 하면, 먼저 가장 뚜렷하게 보이는 것은 normal probability plot에서 정규성이 위배되는 것이다. 다음으로, 오차의 평균은 0에 가까운 것으로 보이며, 예측변수와 반응변수간에 뚜렷한 비선형성은 드러나지 않는다. 등분산성의 경우는 잔차플롯만으로 판단하기 어렵다고 판단하였다. 관측값 수가 너무 많기 때문에, 잔차 플롯 상 몇 개의 점이 잔차가 큰 것을 이분산성의 근거로 볼 수 없다고 생각하였다. 따라서 등분산성 여부를 판단하기 위해 추가적인 데이터 처리작업을 실시하였다. 절차는 다음과 같다.

1. 결정된 다중회귀모형에서 rstudent 잔차 산출,
2. Y의 범위 산출 (univariate 프로시저 이용)
3. Y값이 중복되는 경우는 드물 것이므로, y의 범위를 총 20구간으로 나누어 각각의 그룹마다 rstudent의 분산 산출.

이를 수행한 코드는 다음과 같다.

**proc** **reg** data = stdmarket;

model y = pay busstop apart\_price var10 working\_pop / r;

output out = res rstudent = r;

**run**; **quit**;

**proc** **univariate** data = res;

var y;

**run**;

**data** res2;

set res;

if y < -**1** then ygroup = -**1** ;else if y < -**0.6** then ygroup = **0** ;

else if y < -**0.4** then ygroup = **1**; else if y < -**0.2** then ygroup = **2**;

… (중략)

else if y < **4** then ygroup = **19**; else ygroup = **20**;

**run**;

**proc** **means** data = res2 mean std;

class ygroup;

var r;

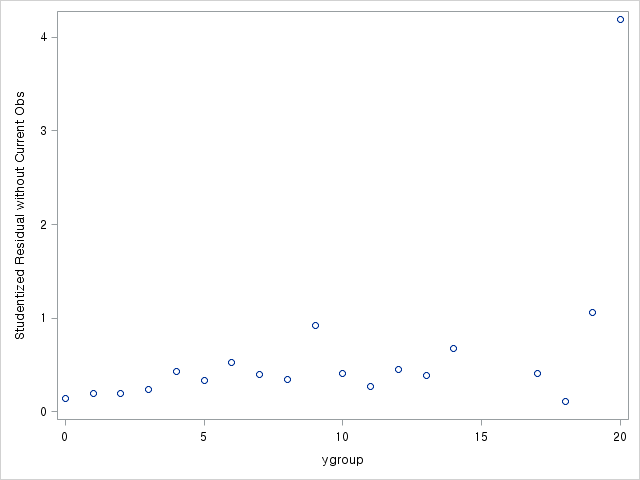
output out = variance std = s;

**run**;

**proc** **sgplot** data = variance;

scatter x = ygroup y = s;

**run**;

이를 통해 얻은 그래프는 다음과 같다.

<그림 5 : y의 변화에 따른 분산의 변화>

그림 5를 보면, y가 커짐에 따라 분산값이 증가한다고 보기 어려움을 알 수 있다. (ygroup = 20에 속한 데이터 들은 그 수가 적으며 분포된 범위가 매우 넓다 ; ygroup의 범위는 0.2이지만 ygroup = 20에는 4보다 큰 모든 데이터들 (개수 < 20) 이 들어있다. Y의 최대값은 18이다. 따라서 ygroup = 20 은 outlier로 간주하고, 분석에서 제외한다)

실제로, 위의 그래프의 플롯에 선형회귀모형을 적합시키면, y가 커질수록 분산은 다소 증가하는 경향이 있으나, 모델이 유의하다고 볼 수 없다는 결과가 나온다. 테이블이(가) 표시된 사진

자동 생성된 설명

<그림 6: 그림 5의 그래프에 선형회귀 적합>

따라서, 우리가 고려하는 다중회귀모형에서, 오차의 등분산성은 만족되는 것으로 판단한다.

결국, 모형타당성 검증에 의하여,

모델 y = B0\*pay+ B1\* busstop+ B2\*apart\_price + B3\*var10 + B4\*working\_pop은 오차의 정규성을 만족하지 않고 있음을 알 수 있다. 오차의 정규성은 box-cox 변환을 통하여 해결 할 수 있다. 다만, box-cox 변환은 y에 대한 power transformation이기 때문에, y가 음수일 때는 계산 할 수 없다. 따라서, 기존의 y값에 1씩 더한 y\_plus라는 새로운 변수를 만든 후 Box-Cox 변환을 실시하겠다.

**proc** **transreg** details data = stdmarket;

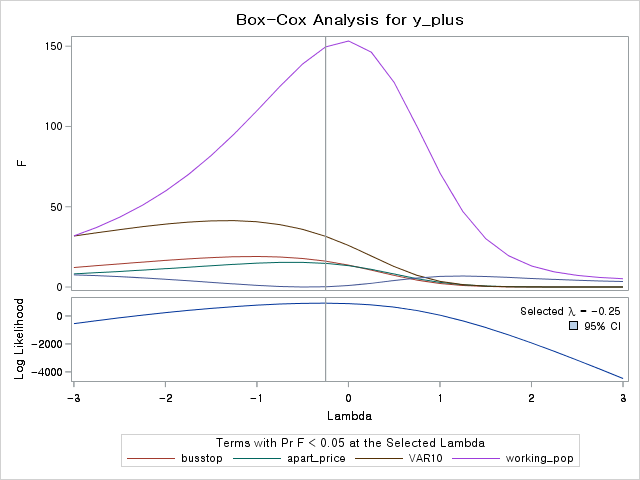
model boxcox(y\_plus) = identity(pay busstop apart\_price var10 working\_pop);

**run**; **quit**;

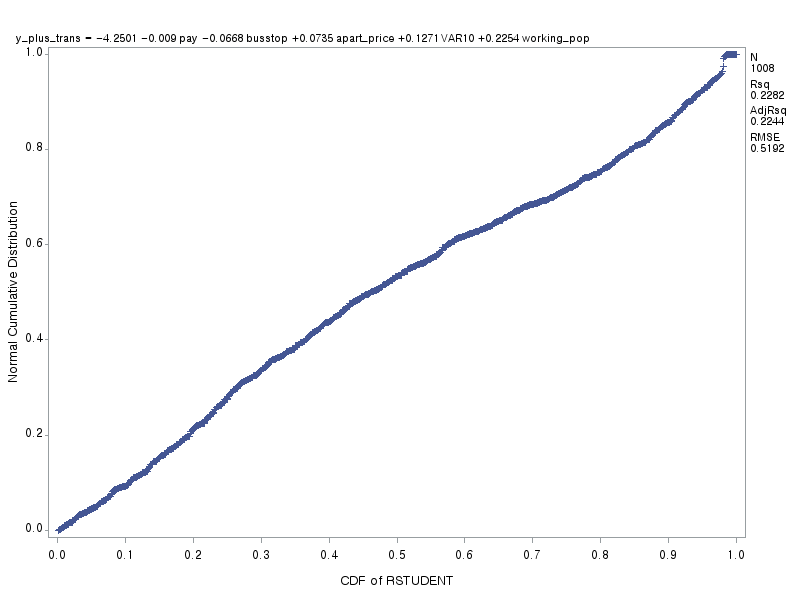
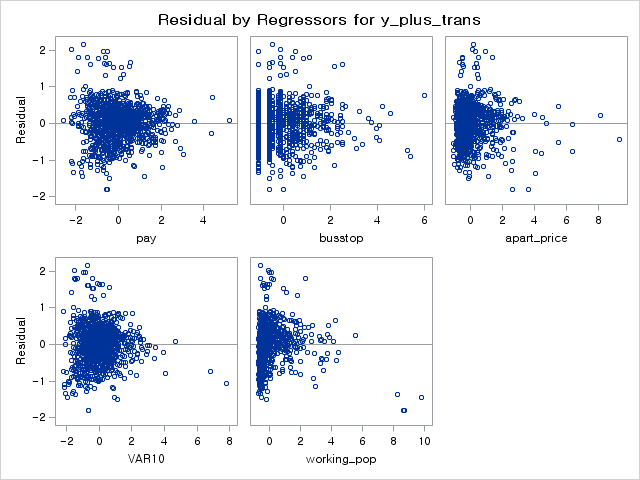
코드는 위와같다. Transreg 프로시저를 사용하였다.

위 과정에서 구한 결과는 lambda = -0.25이고, 이 값을 이용하여, 새로운 변수 y\_plus\_trans를 만들었다.

Y\_plus\_trans = y\_plus^-0.25 / -0.25



<그림 8: box-cox 변환 lamda 값 산출 결과 = 0.25>



<그림 9: box-cox 변환 후 정규성 검정 결과. 정규성을 만족하는 것을 볼 수 있다.>

구해진 y\_plus\_transform 변수를 이용해 새로운 모델을 적합 후, 모형 타당성 검증을 시행하였다.

그림 9에 따르면, box-cox 변환 후 정규성 검정결과, 정규성을 만족하는 것을 볼 수 있다. 다만, box-cox 변환 후 변수들에 대한 잔차플롯을 볼 때, y의 평균이 커질수록 분산이 작아지는 이분산성이 발견된다. 이의 해결을 위해 다시 한번 가중회귀를 시행한다.

가중회귀는 일단 각 y의 수준별 분산 값이 필요하다. 따라서, 앞 부분에서 다룬 것과 같이, y를 적절한 구간을 나누어 구간별로 분산값을 구하고, 이의 역수를 가중치로 두고 가중회귀를 실시한다. 코드는 다음과 같다.

/\*10분위 값 구하기\*/

**proc** **univariate** data = stdmarket;

var y\_plus\_trans ;

output out = work.dat pctlpre = p pctlpts= **0** to **100** by **10**;

**run**;

/\*가중회귀\*/

**proc** **reg** data = stdmarket;

model y\_plus\_trans = pay busstop apart\_price var10 working\_pop;

output out = work.res rstudent = r;

**run**;**quit**;

/\*10분위 값을 바탕으로 10등분\*/

**data** res;

set res;

if y\_plus\_trans < -**5** then y\_adj\_group = **1**; else if y\_plus\_trans < -**4.7** then y\_adj\_group = **2**;

else if y\_plus\_trans < -**4.5** then y\_adj\_group = **3**; else if y\_plus\_trans < -**4.36** then y\_adj\_group = **4**;

else if y\_plus\_trans < -**4.2** then y\_adj\_group = **5**; else if y\_plus\_trans < -**4.1** then y\_adj\_group = **6**;

else if y\_plus\_trans < -**3.95** then y\_adj\_group = **7**; else if y\_plus\_trans < -**3.8** then y\_adj\_group = **8**;

else if y\_plus\_trans < -**3.57** then y\_adj\_group = **9**; else y\_adj\_group = **10**;

**run**;

**proc** **means** std;

class y\_adj\_group;

var r;

output out = res2 std = v;

**run**;

**proc** **sort** data = res;

by y\_adj\_group;

**run**;

**data** res3;

merge res res2;

by y\_adj\_group;

**run**;

**data** res3;

set res3;

w = **1**/v;

**run**;

**proc** **reg** data = res3;

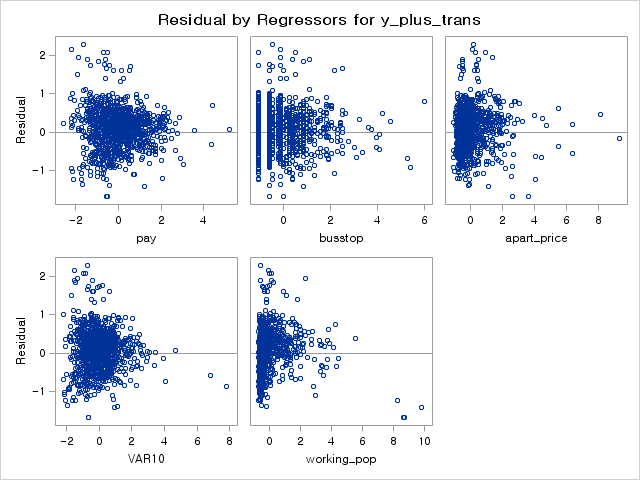
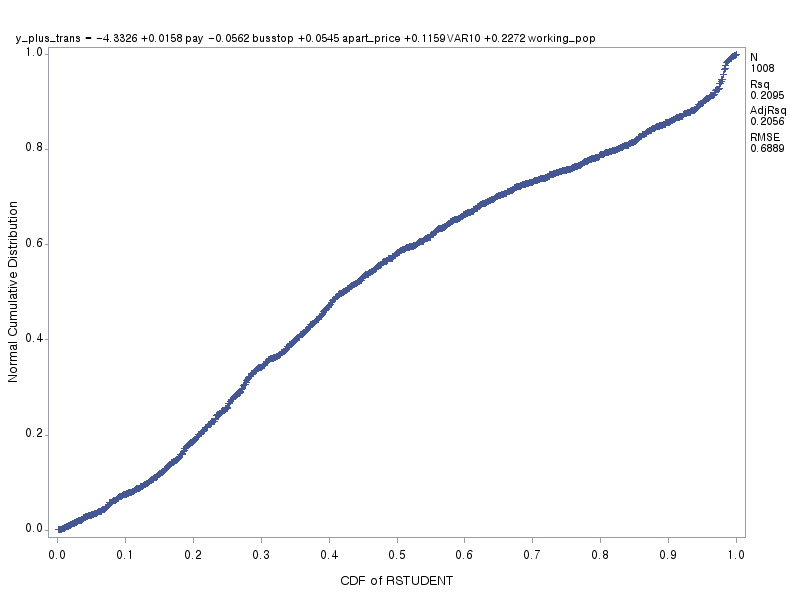
model y\_plus\_trans = pay busstop apart\_price var10 working\_pop;

weight w;

plot npp. \*rstudent.;

**run**; **quit**;

위 프로그램에서 구한 모델로 다시 모형 타당성 검증을 하면 다음과 같다.



<그림 9 ; 모형타당성 검증 결과>

Box-cox 변환만 시행했던 이전 모델에 비하여 다소 정규성이 흐트러진 느낌이 있지만, 심각한 정규성 위배는 발견되지 않는다. 또한, 변수들을 대상으로 한 잔차플롯을 보면, 가중회귀시행 전과 비교했을 때 이분산성이 확연히 해소된 것을 볼 수 있다.

따라서 위 가중회귀모형은 타당하다.

4. 회귀진단

다음으로, 회귀진단을 수행한다. 먼저 잔차를 구한 코드는 다음과 같다.

**proc** **reg** data = res3;

model y\_plus\_trans = pay busstop apart\_price var10 working\_pop / r influence;

weight w;

output out=res4 dffits= dfffits covratio = covratio;

**run**; **quit**;

영향력관측치 분석에서 고려하는 통계량

이 대략 5가지 있는데, Cook’s D, leverage, covratio, dffits, dfbeta가 바로 그것이다.

각 지표가 각 관측치의 influential을 판단하는 기준은 다음 표로 정리한다.

|  |  |
| --- | --- |
| 지표 | 구간 |
| Cook’s D | D > 1 |
| Leverage | X > 0.014 = 2\*p/n |
| Dffits | |X| > 0.167 = 2\*(p/n)^0.5 |
| Covratio | X > 1.042 = 1 + 3p/n or X < 0.958 = 1 – 3p/n |
| Dfbeta | |x| > 2/(n)^0.5 이지만 이 데이터에서는 이 지표는 고려하지 않겠다.  (너무 데이터가 많은 관계로) |

<그림 10 : 위의 코드를 통해 얻어진 값 중 영향력이 있을 것으로 보이는 관측치들>

텍스트, 신문이(가) 표시된 사진

자동 생성된 설명이 값들은 회귀진단을 통해 얻어진 영향력 관측치 중 일부이다. 이들을 어떻게 처리할 것인지가 회귀진단에서의 주요 이슈인데, 이 데이터들이 유효성여부와 데이터 수집과정에서 오류가 있었는지의 여부는 현 시점에서 알 수 없다. 다만, 데이터의 출처가 믿을 만 한 곳 이기에(서울 공공데이터 포털) 관측치는 유효한 값 일 것이라고 예상

한다, 결론적으로, 이 영향력 관측치들을 기존 데이터에서 삭제하지 않고, 계속 포함시키며 모델링 하겠다. 이를 끝으로, 데이터 분석 단계는 종료된다.

테이블이(가) 표시된 사진

자동 생성된 설명**3. 결론**

**3.1. 분석 결과 요약**

옆의 그림은 최종선택한 회귀모형을 적합한 결과이다.

이에 따르면, 적합된 모델은

**Y\_plus\_trans = -4.33 + 0.016pay -0.056busstop +0.054apart\_price + 0.12var10 + 0.23working\_pop**

의 형태를 띈다. 이 모델은 anova 유의성 검정에 의해 유효하며, pay를 제외한 모든 회귀계수들 또한 유의하다. 이 모델을 해석할 때 주의 할 것은, 반응변수가 Y\_plus\_trans, 즉 transformation을 거친 것을 고려 해야 한다는 것이다. 반응변수와 원래 y의 관계는 다음과 같다.

**Y\_plus\_trans = (Y + 1)^-0.25 / -0.25, 즉 Y = (-0.25\*Y\_plus\_trans)^(-4) – 1**

이 관계에 따라,

**(Y + 1)^-0.25 / -0.25 = -4.33 + 0.016pay -0.056busstop +0.054apart\_price + 0.12var10 + 0.23working\_pop**

이고, 즉,

**Y = (1.08 -0.004pay +0.014busstop - 0.0135apart\_price - 0.03var10 -0.58working\_pop)^(-4) -1**

이 된다. 이를 해석하면 다음과 같다.

-pay가 1단위 증가 할 때, 1 / (y+1)^0.25는 -0.004만큼 감소한다. 즉 pay와 y는 양의 상관관계를 가진다.

-busstop가 1단위 증가 할 때, 1 / (y+1)^0.25는 0.014만큼 증가한다. 즉 busstop와 y는 음의 상관관계를 가진다.

-apart\_price가 1단위 증가 할 때, 1 / (y+1)^0.25는 -0.0135만큼 감소한다. 즉 apart\_price와 y는 양의 상관관계를 가진다.

-var10가 1단위 증가 할 때, 1 / (y+1)^0.25는 -0.03만큼 감소한다. 즉 var10와 y는 양의 상관관계를 가진다.

-working\_pop가 1단위 증가 할 때, 1 / (y+1)^0.25는 -0.58만큼 감소한다. 즉 working\_pop와 y는 양의 상관관계를 가진다.

-한 단위가 변 할 때 y에 대한 효과가 가장 큰 것은 working\_pop이다. 다른 변수보다 working\_pop이 타 지역에 비해 높은 지역이 비교적 매출이 잘 나올 것이다.

- pay의 경우, 회귀계수 검정에 의하면 Y와 상관관계가 없다는 귀무가설을 기각하지 못함을 알 수 있다.

요약하면, 아파트 가격, 지역 생활인구 수, 지역 직장인구 수는 지역 골목상권 매출과 유의미한 양의 상관관계가 발견된다. 반면 주변에 버스정류장 수는 지역 골목상권 매출과 음의 상관관계가 발견되었다. 이는 우리의 직관에 반하는 결과인데, 만약 실제 관계에서 버스정류장 수와 상권 매출이 양의 상관관계가 있다고 가정하면, 우리의 모델 회귀계수가 이와 반대로 나온 것으로 추측되는 이유는 다음의 몇 가지가 있다.

1. 회귀계수의 분산이 너무 큼
2. 중요한 설명변수가 모델에 포함되지 않음.
3. 다중공선성
4. 타 변수들 간 관계

우선, 다중공선성은 제거 되었으므로 3은 회귀계수 부호 오류의 이유가 될 수 없다.

4의 경우, 버스정류장이 많은 지역에서 Y와 양의 관계를 가지는 설명변수들의 값들이 작은 가 생각해보면, 직관적으로 그렇지 않다고 볼 수 있다. (직관적으로 버스정류장이 많은 지역에서 아파트 가격, 직장인구, 생활인구가 적을 일은 없을 것이다.) 따라서, 4 또한 부호 오류의 원인이라고 보지 않는다.

가장 유력한 원인은 2인데, 우선, 본 연구에서 구해진 모델의 결정계수(R-squared)는 0.2 가량으로 낮다. 즉, Y를 설명하는 중요한 변수가 설명변수에서 빠져 있다는 것을 알 수 있다. 이는 본 연구의 한계점이라고도 볼 수 있다.

**3.2. 분석의 장점 및 한계점 설명**

본 분석의 장점은, 신뢰할 만한 데이터를 바탕으로 하여, 모형타당성을 최대한 만족시키고자 가중회귀, Box-cox 변환 등을 거쳐 최대한 엄밀하게 선형모형을 적합하였다는 것이다. 반면, 본 분석의 한계점으로는, 첫째, 데이터의 출처는 신뢰할 만 하였으나 사용된 설명변수가 Y의 변동의 상당부분을 설명하기에는 부족했다는 점. 즉, 중요한 설명변수가 하나 이상 고려되지 못했다. 이는 모델의 낮은 R-Squared 통계량에서 드러난다. 둘째, 본 모델은 해석이 어렵다는 단점이 있다. 특별히 변환되지 않는 단순선형모형과 달리, 이 모델은 box-cox 변환으로 인하여 구해진 모델을 다시 y에 대해 변환시켜주는 과정에서 직관적인 해석이 어렵게 되었다. 그 다음으로, 잔차분석을 할 때, 특히 선형성이나 오차의 합이 0임을 추론하는 과정에서, 정확한 측정절차나 통계적 검정 없이 시각적인 직관을 이용하여 모형타당성을 검증하였다. 관측값의 수가 적다면 상관이 없겠으나, 대략 1000개 정도의 관측값을 가지고 있는 본 모델의 경우 적절하지 않는 잔차플롯 분석 과정이다. 앞으로 이루어지는 분석에서는, 이러한 한계점들을 보완하여 데이터 분석을 할 필요성이 있겠다.

* 1. **추가 연구사항 제안**

본 연구를 시작할 때, 본 보고서에 드러난 것 외에 도출해보고 싶은 결론이 더 있었다. 첫째로, 국가에서 제공한 지역별 공시지가 데이터를 추가적으로 참고하여, apart\_price 대신에 설명변수로 추가하여, (아파트 가격은 그 지역 거주민들의 부의 수준은 나타낼 수 있으나 가게 면적 당 임대료 등을 암시할 수 있는 지표는 아니다.) 땅값과 매출간 상관관계 분석과 더불어, 과연 면적 당 임대료가 비싼 만큼 매출에 효과가 있는지, 즉 비싼 돈 내고 점포를 운영 할 만 한지에 대한 분석을 해보고 싶었으나, 추가적인 데이터사용에 대한 문제와, 경제학적 지식의 부족, 더 자세한 정보가 필요하다는 점 때문에 실행에 옮기지는 못하였다. 위 조건들이 갖추어진다면, 또는 이를 실행에 옮기기 위한 창의적인 방법이 떠오른다면, 이에 대한 구체적인 연구를 해보는 것도 흥미로울 것이라고 생각한다.

다음으로, 선형모델 외 새로운 모델로 같은 데이터분석을 실시하였을 때 나오는 결과값들에 대하여도 연구해 볼 수 있겠다. 그 중에서도, Clustering을 활용하여 특성이 비슷한 지역들 끼리 군집화 하는 것도 활용도 높은 결과를 창출 할 수 있을 것으로 기대한다.

**# References**

**Douglas C.Montgomery, Introducction to Linear Regression Analysis. 5th ed.**

**# Appendix : SAS CODE**

/\*데이터 불러오기\*/

**proc** **import** file = "C:\Users\jmjwj\Documents\UOS\2학년 2학기\회귀분석\market2021.csv"

dbms = csv out = work.market;

**run**;

/\*데이터 표준화\*/

**proc** **standard** data = market out = stdmarket mean = **0** std= **1**;

**run**;

/\* 모형선택 \*/

**proc** **reg** data = stdmarket;

model y = income pay subway busstop school apart\_count apart\_price var10 working\_pop / vif collin;

**run**; **quit**;

/\*별다른 다중공선성 발견되지 않음\*/

/\*backward selection - 프로그램에서는 모든 계수 p-value가 0.1보다 작으면 멈춘다.\*/

**proc** **reg** data = stdmarket;

model y = income pay subway busstop school apart\_count apart\_price var10 working\_pop / selection = backward;

**run**; **quit**;

/\*cp 가 최소인 모델을 결정\*/

**proc** **reg** data = stdmarket;

model y = pay busstop apart\_price var10 working\_pop;

**run**; **quit**;

/\*모형 타당성 검증\*/

**proc** **reg** data = stdmarket;

model y = pay busstop apart\_price var10 working\_pop / partial;

plot rstudent. \*(predicted. pay apart\_price busstop var10 working\_pop);

plot npp. \*rstudent.;

**run**; **quit**;

/\*box-cox 시도 : y가 양수여야지만 가능\*/

**proc** **transreg** details data = market;

model boxcox(y) = identity(pay busstop apart\_price var10 working\_pop);

**run**; **quit**;

/\* 정규성 위배. 평균은 0으로 보임. 등분산성은 애매함. 독립성 고려하지 않을거고 선형성은 있는 듯.\*/

/\* 등분산성 확인해보자\*/

**proc** **reg** data = stdmarket;

model y = pay busstop apart\_price var10 working\_pop / r;

output out = res rstudent = r;

**run**; **quit**;

**proc** **univariate** data = res;

var y;

**run**;

**data** res2;

set res;

if y < -**1** then ygroup = -**1** ;else if y < -**0.6** then ygroup = **0** ;

else if y < -**0.4** then ygroup = **1**; else if y < -**0.2** then ygroup = **2**;

else if y < **0** then ygroup = **3**; else if y < **0.2** then ygroup = **4**;

else if y < **0.4** then ygroup = **5**; else if y < **0.6** then ygroup = **6**;

else if y < **0.8** then ygroup = **7**; else if y < **1** then ygroup = **8**;

else if y < **1.2** then ygroup = **9**; else if y < **1.4** then ygroup = **10**;

else if y < **1.6** then ygroup = **11**; else if y < **1.8** then ygroup = **12**;

else if y < **2** then ygroup = **13**; else if y < **2.2** then ygroup = **14**;

else if y < **2.4** then ygroup = **15**; else if y < **2.6** then group = **16**;

else if y < **3** then ygroup = **17**; else if y < **3.5** then ygroup = **18**;

else if y < **4** then ygroup = **19**; else ygroup = **20**;

**run**;

**proc** **means** data = res2 mean std;

class ygroup;

var r;

output out = variance std = s;

**run**;

**proc** **sgplot** data = variance;

scatter x = ygroup y = s;

**run**;

**proc** **reg** data = variance;

model s = ygroup;

**run**;

**proc** **reg** data = trans\_data;

model y = pay busstop apart\_price var10 working\_pop;

weight weight;

plot npp. \*rstudent.;

plot rstudent. \*(predicted. pay busstop apart\_price var10 working\_pop);

**run**; **quit**;

/\* 더 발전된 결과 \*/

/\*box-cox \*/

**data** stdmarket;

set stdmarket;

y\_plus = y + **1**;

**run**;

**proc** **transreg** details data = stdmarket;

model boxcox(y\_plus) = identity(pay busstop apart\_price var10 working\_pop);

**run**; **quit**;

**data** stdmarket;

set stdmarket;

y\_plus\_trans = y\_plus\*\*(-**0.25**)/ -**0.25**;

**run**;

**proc** **reg** data = stdmarket;

model y\_plus\_trans = pay busstop apart\_price var10 working\_pop;

plot npp. \*rstudent.;

plot rstudent. \*(predicted. pay busstop apart\_price var10 working\_pop);

**run**;**quit**;

**proc** **univariate** data = stdmarket;

var y\_plus\_trans ;

output out = work.dat pctlpre = p pctlpts= **0** to **100** by **10**;

**run**;

/\*가중회귀 \*/

**proc** **reg** data = stdmarket;

model y\_plus\_trans = pay busstop apart\_price var10 working\_pop;

output out = work.res rstudent = r;

**run**;**quit**;

**data** res;

set res;

if y\_plus\_trans < -**5** then y\_adj\_group = **1**; else if y\_plus\_trans < -**4.7** then y\_adj\_group = **2**;

else if y\_plus\_trans < -**4.5** then y\_adj\_group = **3**; else if y\_plus\_trans < -**4.36** then y\_adj\_group = **4**;

else if y\_plus\_trans < -**4.2** then y\_adj\_group = **5**; else if y\_plus\_trans < -**4.1** then y\_adj\_group = **6**;

else if y\_plus\_trans < -**3.95** then y\_adj\_group = **7**; else if y\_plus\_trans < -**3.8** then y\_adj\_group = **8**;

else if y\_plus\_trans < -**3.57** then y\_adj\_group = **9**; else y\_adj\_group = **10**;

**run**;

**proc** **means** std;

class y\_adj\_group;

var r;

output out = res2 std = v;

**run**;

**proc** **sort** data = res;

by y\_adj\_group;

**run**;

**data** res3;

merge res res2;

by y\_adj\_group;

**run**;

**data** res3;

set res3;

w = **1**/v;

**run**;

**proc** **reg** data = res3;

model y\_plus\_trans = pay busstop apart\_price var10 working\_pop;

weight w;

**run**; **quit**;

/\*회귀진단\*/

**proc** **reg** data = res3;

model y\_plus\_trans = pay busstop apart\_price var10 working\_pop / r influence;

weight w;

output out=res4 dffits= dffits covratio = covratio h= leverage;

**run**; **quit**;

**data** res4;

set res4;

dffits\_abs = abs(dffits);

**run**;

**proc** **sort** data = res4 ;

by descending dffits\_abs leverage ;

**run**;

1. 제프리 울드리지(1960~). Introductory Econometrics: A Modern Approach [↑](#footnote-ref-1)