[다변량 통계학 최종보고서]

- 스타벅스 매장 입점 장소 판별 및 분류 -

서울시립대학교

다변량 통계학(김규성 교수님)

2016580009 통계학과 김태현

차례

---

1. 서론

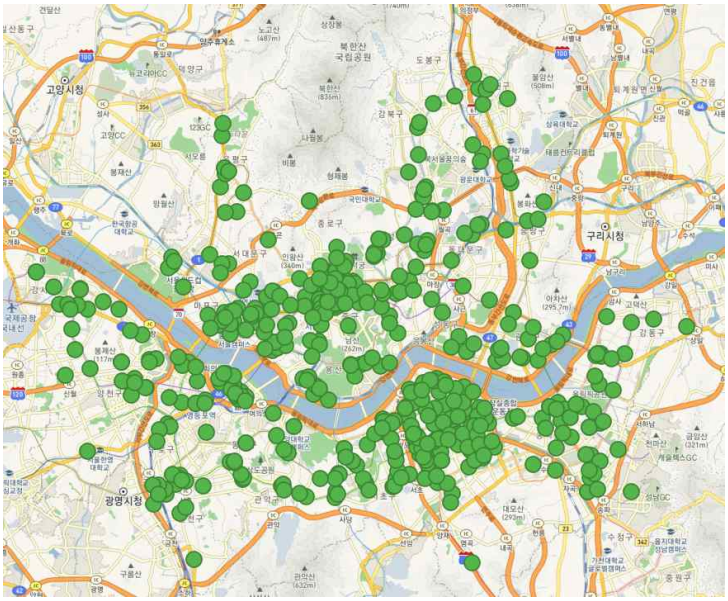
1.1 연구 목적

스타벅스는 1999년 대한민국에서 최초로 이화여대점을 시작으로, 성장을 거듭하며 2020년 기준 약 1,300개 이상의 점포를 운영 중이다. 서울의 경우 약 520개의 매장이 있으며 강남구는 매장이 80개로 구 중에서 가장 많은 스타벅스 매장이 입점해있고 가장 적은 구는 2개의 매장뿐인 도봉구이다. 주요 상권이 있는 구를 제외하면 평균적으로 구별로 10개 정도의 매장이 있다. 스타벅스는 가맹점을 내지 않고 100% 직영점으로만 운영하기 때문에 스타벅스가 입점하는 장소에는 분명한 요인이 존재한다.

스타벅스는 고급화 전략을 추구한다. 유동인구가 많은 지역의 랜드마크 건물에 넓고 쾌적한 공간을 가진 카페형 매장을 설계하고 무료 와이파이와 편안한 음악을 제공함으로써 고객들은 장시간 매장에 머물면서 커피를 즐길 수 있게 한다. 이 전략은 적중했고 스타벅스라는 브랜드가 가지는 이미지만으로 많은 사람을 끌어들인다.

따라서, 스타벅스가 어떤 상권에 입점하게 되면 주변 카페의 매상을 흡수하는 경향이 있다. 그러므로 새롭게 카페를 개점할 때 스타벅스가 주변에 매장을 낼 것인지에 대한 예측은 신규 창업자에게 매우 중요한 요소이다. 특히 최근 스타벅스는 배달 서비스를 시작하여 스타벅스의 입점이 주변의 업체에 미치는 영향력이 더 넓어졌으므로 입지 선정에서의 중요도가 더 올라갔다고 할 수 있다.

본 연구에서는 판별 및 분류 통계 기법을 이용하여 동별로 스타벅스의 입점 여부를 반응변수로 하고 유동인구, 임대료 가격, 인접한 다른 매장과의 거리 등 여러 요인을 설명변수로 하여 스타벅스가 입점할만한 장소와 그렇지 않은 장소를 분류한다. 카페 또는 유사한 업종을 새롭게 창업하려는 사람들에게 입지선정 단계에서 고려할만한 지표를 주는 것이 본 연구의 목적이다.



서울시 스타벅스 매장의 위치

1.2 문헌 연구

"서울시 프랜차이즈 커피점 입지의 영향요인, 스타벅스 커피전문점을 중심으로, 정승영, 최인섭"을 참고하였다. 이 연구는 다중 회귀분석을 사용하여 스타벅스의 위치에 영향을 미치는 요인을 탐색하는 것이다. 위치 이론에 기반을 두고 있으며 스타벅스의 입점 요인을 상가보증금, 월세, 상가권리금, 총사업체 수, 3차 산업체 수, 3

차 산업 종사자 수, 제조업체 수, 인구밀도로 하여 분석하였다. 그 결과 상가 월세, 상가보증금, 상권의 인구 및 산업이 커피전문점의 입지에 영향을 주는 중요한 변수임을 제시하였다.

"위치정보를 활용한 커피 전문점의 입점 분석, 이동엽, 윤영태"이라는 연구가 있다. 해당 연구에서는 서울 시내 유명 커피 프랜차이즈 매장의 위치 정보를 수집하고 지리통계를 이용하여 입점 전략을 실증적으로 분석하였다. 단순히 매장이 속한 지역의 분류 정보만을 이용한 것이 아니라 위도와 경도로 표현한 실제 위치를 추출하여 매장 사이의 거리를 측정하고 서비스 전략을 실증적으로 추정하였다. 결론에 따르면 고급화 전략을 추구하는 스타벅스는 사업성이 높은 지역에는 다수의 매장을 개설하는 집중 초토화 전략을 구사하고 있다는 것을 확인하였고, 저가 전략의 이디야 커피는 좋은 상권에서는 스타벅스 옆자리를 차지하여 고객을 확보하는 동시에 스타벅스가 없는 지역에서는 잠재 고객의 분포에 따라 매장을 늘려간다는 것을 알 수 있었다. 해당 연구에서 아쉬운 점으로 언급한 것은 지하철역 출구의 위치 정보를 이용하지 못한 것인데 본 연구에서 다뤄볼 만한 가치가 있다고 판단된다.

1.3 데이터 설명

서울시에는 25개 자치구와 425개 행정동이 있다. 사용한 데이터는 상가정보 데이터, 지하철역 위치 데이터, 유동 인구 데이터, 지가 데이터이다. 모든 데이터는 2019년을 기준으로 작성되었다. 데이터는 425개의 행과 11개의 열로 이루어져 있다. 10개의 변수는 스타벅스 입점 여부(starbucks), 유동인구(pop0 ~ pop70), 지하철역과의 거리(subway) 그리고 지가(landprice)이다. index는 행정동(dong)이다.

상가정보 데이터는 공공 데이터 포털이 출처의 소상공인시장진흥공단_상가(상권) 데이터를 사용하였다. 서울시의 모든 상가에 대한 데이터가 존재한다. 상호명, 업종, 시군구, 행정동, 주소, 위도, 경도 등 상가에 대한 모든 기본정보를 포함한다. 분석의 기본 틀이 되는 데이터로 특정 동에 스타벅스가 입점해있는지 여부, 지하철과의 거리 등에 대한 정보를 얻을 수 있다. 이 데이터에서는 스타벅스 입점여부(starbucks) 변수를 추출했다.

지하철역 위치 데이터는 서울 열린데이터광장 출처의 지하철역정보 데이터를 사용하였다. 서울시의 지하철의 첫차, 막차, 노선, 위치정보를 얻을 수 있으며 분석에서는 지하철역의 위치정보만을 사용한다. QGIS를 이용해서 상가정보 데이터와 지하철 위치 데이터를 연계하여 각 동의 상가들과 지하철 사이의 평균 거리(subway)변수를 얻었다.

유동 인구 데이터는 서울특별시 빅데이터 캠퍼스의 서울시 행정동단위 월별 KT 유동 인구 데이터를 사용하였다. 월별 행정동 단위로 연령별 유동인구 수를 집계한 데이터이다. 연령은 10세 이하부터 70세 이상까지 10세 단위로 구분되어 있다. 해당 데이터에서는 유동인구(pop0~pop70)변수를 얻었다. pop10부터 pop60은 10세 단위의 연령대를 의미하며 pop0은 10세이하, pop70은 70세 이상이다.

지가 데이터는 한국 감정원 부동산 통계 정보 시스템의 지가 변동 데이터를 사용하였다. 서울시 행정동 단위의 지가를 나타내는 상대적인 지표이다. 스타벅스는 모든 매장을 임대 형식으로 운영하기 때문에 임대료 데이터를 사용하는 것이 더 적절하겠지만 임대료 데이터는 행정동별로 집계된 데이터가 존재하지 않았다. 또한 지가가 개별 상가의 임대료 정보를 나타낼 순 없겠지만 행정동별로 평균적인 임대료를 행정동의 지가지수가 대표할 수 있다는 생각에 기반하여 분석에 사용할 데이터로 채택하였다. 지가(landprice) 변수를 얻을 수 있었다.

1.4 분석 방법

본 연구는 SAS와 R, QGIS를 사용하여 분석을 진행한다. 대부분의 분석 과정은 SAS를 사용하여 진행할 것이고 R과 QGIS는 데이터를 처리하는 데에 주로 사용할 것이다.

판별(Discrimination) 및 분류(Classification) 분석은 집단에 대한 정보로부터 집단을 구별할 수 있는 판별함수 또는 판별규칙을 만들고, 새로운 개체에 대해 어느 집단에 속하는지를 판별하여 분류하는 다변량 기법으로 집단에 대한 정보를 이용한 탐색적인 통계 기법이다.

서울의 모든 행정동을 스타벅스의 입점 여부에 따라 두 집단으로 나누어 판별 및 분류 분석을 진행할 것이다. 스타벅스가 입점한 장소의 정보를 바탕으로 두 집단을 가장 잘 구별할 수 있는 판별함수(판별규칙)을 만들고 그에 따라서 어떤 특정 장소(동)가 스타벅스가 입점할만한 장소인지 그렇지 않은 장소인지 분류한다.

두 집단을 판별하는 방법에는 크게 Fisher 방법과 우도함수를 이용한 방법이 있다. 우선 피셔방법은 거리를 이용한 판별방법으로 정규성 가정은 필요하지 않다. 두 그룹의 평균 차이를 최대화하는 함수를 찾고, 그 함수(Y)는 다음과 같이 표현할 수 있다.

$$l : \max_{l(\neq 0)} \left(\frac{\mu_{1Y} - \mu_{2Y}}{\sigma_Y} \right)^2 = \frac{(l'\delta)^2}{l'\Sigma l} = \delta' \Sigma^{-1} \delta \Leftrightarrow l = c\Sigma^{-1}\delta = c\Sigma^{-1}(\mu_1 - \mu_2)$$

$$Y = l'X,\ \delta = \mu_1 - \mu_2$$

우도함수를 이용한 방법은 확률을 이용한 판별방법으로 정규성 가정이 필요하다. 우도함수 방법은 다음과 같은 오분류 확률을 최소화하는 최적 분류규칙을 사용한다. 공분산 행렬이 다른 경우이다.

$p_1 : X$가 G_1에서 발생할 사전 확률, $p_2 : X$가 G_2에서 발생할 사전 확률,

$f(X|G_i) = (2\pi)^{-p/2} |\Sigma_i|^{-1/2} e^{-(X-\mu_i)'\Sigma_i^{-1}(X-\mu_i)/2}$,

$Q(X) = \ln \frac{f(X|G_1)}{f(X|G_2)} = \frac{1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) - \frac{1}{2}(X-\mu_1)'\Sigma_1^{-1}(X-\mu_1) + \frac{1}{2}(X-\mu_2)'\Sigma_2^{-1}(X-\mu_2)$라고

하면, $Q(X) > \ln(\frac{p_1}{p_2})$이면 X을 G_1그룹에, 그렇지 않으면 G_2그룹에 분류한다.

분류함수의 능력을 판단하기 위해 재대입 분류, 교차 타당성 등을 이용하여 오류율을 계산한다. 재대입법은 데이터로부터 유도된 판별함수를 다시 데이터에 적용하여 오분류 되는 표본의 수를 전체 표본의 수로 나누어 오분류율을 계산한다. 교차 타당성에 의한 방법은 한 개만의 표본을 제외한 나머지 표본으로 판별함수를 계산하고 구해진 판별함수를 이용해 제외되었던 표본을 분류한다. 이와 같은 과정을 전체 표본의 크기만큼 시행하여 오분류율을 계산한다.

1.5 결과 활용 및 기대 효과

분석 결과로 각 연령별 유동 인구, 상가들의 지하철과의 거리, 지가지수를 기반으로 한 판별함수를 얻을 수 있다. 특정 동에 대한 정보를 판별함수에 대입하면 스타벅스가 입점할지에 대한 여부를 측정할 수 있게 된다. 즉, 새롭게 카페 및 디저트 창업을 고려하고 있는 사람들은 자신이 고려하고 있는 입지가 스타벅스가 향후에 입점할 만한 장소인지 확인할 수 있다. 이는 입지 선정에 고려할만한 중요한 지표로 활용될 수 있다.

2. 본론

2.1 분석 방법 소개

우선 기초통계 분석으로 평균, 분산같은 변수들의 특징을 전체적으로 구하고 그룹별로 나누어서 구하여 둘을 비교한다. 그 후 사전 확률을 계산하여 판별에 필요한

p_1, p_2를 구한다.

동일성 검정을 통해 두 그룹의 공분산 행렬이 다른지 확인하고 그 결과 스타벅스가 입점한 동의 그룹과 입점하지 않은 동의 그룹의 공분산 행렬은 다르다는 결론을 내렸다. 따라서 공분산 행렬이 다를 때의 분류 방법을 사용하여 이차형식의 판별함수를 구하였다.

다변량 정규분포를 가정한 모수적 방법을 사용한 모델과 비모수적 방법인 피셔 방법을 사용한 모델. 모든 변수(연령별 유동인구, 지하철과의 거리, 지가지수)를 포함한 모형과 Stepwise 변수 선택법을 통해 선택된 변수를 사용한 모델을 비교한다.

총 4가지의 모형을 재대입법과 교차 타당성을 이용해서 가장 오분류율이 적은 모형을 채택하였다. 마지막으로 새로운 데이터를 선택된 모형의 판별규칙(판별함수)에 대입하여 결과를 확인한다.

2.2 데이터 분석 및 결과 설명

변수	N	평균	표준편차	최솟값	최댓값
starbucks	413	0.4891041	0.5004875	0	1.0000000
X	413	955341.69	7633.35	938573.88	971264.04
Y	413	1950204.24	5732.94	1938159.00	1964649.00
subway	413	449.3399715	331.1582580	37.5055599	2806.73
pop0	413	1138.16	456.6263785	95.9565714	3459.53
pop10	413	1485.87	659.6475192	101.7244000	4471.87
pop20	413	2401.99	1440.45	192.5092451	9599.52
pop30	413	2412.40	1196.07	216.4840000	8009.83
pop40	413	2265.10	955.0921199	197.5745000	6134.53
pop50	413	2146.16	798.3638561	181.2073000	4823.32
pop60	413	1982.33	715.7215756	178.2265000	4501.63
pop70	413	2017.43	760.0427368	183.1454000	4269.55
poptotal	413	15849.43	6153.46	1381.71	35478.80
landprice	413	100.3833060	0.0961180	100.1030000	100.7790000

변수들의 기초통계량이다.

유동 인구 변수를 봤을 때 20, 30, 40, 50대의 유동 인구가 다른 연령대에 비해서 많은 편이다. 10세 이하의 유동 인구의 산포가 제일 작으며 20대의 산포가 제일 크다. 유동 인구, 지가, 지하철과의 거리변수 간에 단위가 달라서 표준화를 해주어야 한다. 특히, 지가(landprice) 변수의 경우 다른 변수와의 단위 차이도 크고 산포도

작아서 표준화가 필수적으로 보인다.

starbucks = 1					
Variable	N	Sum	Mean	Variance	Standard Deviation
pop0	202	232278	1150	210181	458.4547
pop10	202	315462	1562	539993	734.8422
pop20	202	589085	2916	2914718	1707
pop30	202	571652	2830	1884355	1373
pop40	202	514869	2549	1154856	1075
pop50	202	467631	2315	725205	851.5900
pop60	202	422477	2091	565614	752.0732
pop70	202	428837	2123	603487	776.8441
subway	202	78333	387.78727	81120	284.8153
landprice	202	20280	100.39839	0.00911	0.0954

starbucks = 0					
Variable	N	Sum	Mean	Variance	Standard Deviation
pop0	211	237781	1127	207640	455.6753
pop10	211	298202	1413	326018	570.9800
pop20	211	402938	1910	782974	884.8583
pop30	211	424668	2013	674789	821.4554
pop40	211	420616	1993	532687	729.8542
pop50	211	418731	1985	502683	709.0016
pop60	211	396226	1878	441200	664.2293
pop70	211	404363	1916	534736	731.2567
subway	211	107244	508.26720	130377	361.0779
landprice	211	21178	100.36887	0.00898	0.0947

스타벅스가 입점한 그룹(starbucks=1)과 입점하지 않은 그룹(starbucks=0)의 각 변수들에 대한 기초통계량들이다. 평균을 비교해봤을 때, 스타벅스가 입점한 그룹이 모든 연령대에서 입점하지 않은 그룹보다 더 많은 유동 인구(pop)를 보인다. 특히 20, 30 ,40, 50대의 유동 인구에서 더 큰 차이를 보인다. 지하철과의 거리(subway) 또한 입점한 그룹의 평균이 387, 입점하지 않은 그룹의 평균은 508로 큰 차이를 보

인다. 지가(landprice)는 큰 차이를 보이지 않는다.

Class Level Information					
starbucks	Variable Name	Frequency	Weight	Proportion	Prior Probability
0	_0	211	211.0000	0.510896	0.510896
1	_1	202	202.0000	0.489104	0.489104

사전 확률의 경우, 스타벅스가 입점하지 않은 동은 211개로 그 비율이 약 51%, 입점한 동은 202개로 비율이 약 49%로 스타벅스가 입점한 동과 입점하지 않은 동의 사전 확률이 유사하다.

표준화한 변수들의 공분산 행렬이다.

starbucks = 0, DF = 210								
Variable	pop10	pop20	pop40	pop50	pop60	pop70	subway	landprice
pop10	0.749234717	0.311137050	0.528739432	0.560108207	0.574091779	0.581952600	-0.246536414	0.027156400
pop20	0.311137050	0.377358216	0.357962215	0.406896000	0.422387397	0.427868517	-0.221344121	0.024853930
pop40	0.528739432	0.357962215	0.583958125	0.648492304	0.660303728	0.654551231	-0.351639067	0.056760224
pop50	0.560108207	0.406896000	0.648492304	0.788665169	0.804237732	0.798473264	-0.414289122	0.078967655
pop60	0.574091779	0.422387397	0.660303728	0.804237732	0.861286766	0.857855985	-0.446431623	0.085446898
pop70	0.581952600	0.427868517	0.654551231	0.798473264	0.857855985	0.925686040	-0.495254278	0.042928720
subway	-0.246536414	-0.221344121	-0.351639067	-0.414289122	-0.446431623	-0.495254278	1.188860017	-0.040149641
landprice	0.027156400	0.024853930	0.056760224	0.078967655	0.085446898	0.042928720	-0.040149641	0.971699866

스타벅스가 입점하지 않은 그룹의 공분산 행렬이며 S_1에 해당한다.

starbucks = 1, DF = 201								
Variable	pop10	pop20	pop40	pop50	pop60	pop70	subway	landprice
pop10	1.240978621	0.766370527	0.746306022	0.799507283	0.810775483	0.747813143	-0.296686695	-0.065002389
pop20	0.766370527	1.404762153	0.960818959	0.873031770	0.797672529	0.762793958	-0.419726790	0.022105944
pop40	0.746306022	0.960818959	1.266010508	1.103642887	0.979084290	0.912163190	-0.418537064	0.065565119
pop50	0.799507283	0.873031770	1.103642887	1.137782764	1.076677670	1.003496284	-0.428200789	0.084346091
pop60	0.810775483	0.797672529	0.979084290	1.076677670	1.104159897	1.031434445	-0.405013740	0.071345298
pop70	0.747813143	0.762793958	0.912163190	1.003496284	1.031434445	1.044700406	-0.407210868	-0.001028787
subway	-0.296686695	-0.419726790	-0.418537064	-0.428200789	-0.405013740	-0.407210868	0.739699857	-0.036623610
landprice	-0.065002389	0.022105944	0.065565119	0.084346091	0.071345298	-0.001028787	-0.036623610	0.986112365

스타벅스가 입점한 그룹의 공분산 행렬이며 S_2에 해당한다.

Chi-Square	DF	Pr > ChiSq
356.037846	36	<.0001

변수들의 공분산 행렬의 동일성 검정으로 스타벅스가 입점하지 않은 그룹의 공분산 행렬 S_1과 입점한 그룹의 공분산 행렬 S_2가 동일한지 검정한다. 귀무가설은 H_0 : "두 그룹의 공분산이 동일하다."이다. 검정통계량은 약 356.038이며 p-value는 <.0001이다. 따라서 유의수준 5%에서 귀무가설을 기각하게 된다.

즉, 두 그룹은 다른 공분산 행렬을 가지며 합동 공분산 행렬을 사용하는 방법이 아닌 공분산 행렬이 다른 경우에 사용하는 방법인 이차판별함수 형태의 분류규칙을 구해야 한다.

총 네 가지 방법을 통해 분류규칙 및 함수를 구하고 그에 따른 오분류율을 기준으로 모형을 선택할 것이다.

첫째로, 모든 변수를 포함하는 다변량 정규분포를 가정한 모수적 방법이다.

$$X = (X_{pop0}, \cdots, X_{pop70}, X_{landprice}, X_{subway})$$

두 그룹의 공분산 행렬이 같지 않다고 결론 내려졌으므로 그에 따른 우도함수를 이용하는 방법의 분류규칙은 다음과 같다.

$$\begin{aligned} Q(X) &= \ln \frac{f(X|G_1)}{f(X|G_2)} \\ &= \frac{1}{2}\ln\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right) - \frac{1}{2}(\mu_1' \Sigma_1^{-1}\mu_1 - \mu_2\Sigma_2^{-1}\mu_2) \\ &+ (\mu_1'\Sigma_1^{-1} - \mu_2'\Sigma_2^{-1})X - \frac{1}{2}X'(\Sigma_1^{-1} - \Sigma_2^{-1})X \end{aligned}$$

$Q(X) > \ln\left(\frac{p_2}{p_1}\right) \simeq 0 \ (p_1 \simeq p2)$이면 X를 그룹 G_1에, 그렇지 않으면 그룹 G_2에 분류한다.

Number of Observations and Percent Classified into starbucks			
From starbucks	0	1	Total
0	175 82.94	36 17.06	211 100.00
1	96 47.52	106 52.48	202 100.00
Total	271 65.62	142 34.38	413 100.00
Priors	0.5109	0.4891	

Error Count Estimates for starbucks			
	0	1	Total
Rate	0.1706	0.4752	0.3196
Priors	0.5109	0.4891	

재대입법 : 0.3196

Number of Observations and Percent Classified into starbucks			
From starbucks	0	1	Total
0	167 79.15	44 20.85	211 100.00
1	100 49.50	102 50.50	202 100.00
Total	267 64.65	146 35.35	413 100.00
Priors	0.5109	0.4891	

Error Count Estimates for starbucks			
	0	1	Total
Rate	0.2085	0.4950	0.3487
Priors	0.5109	0.4891	

Cross-validation : 0.3487

재대입법 분류에 의한 오류율은 0.3196이고 교차 타당성에 의한 오류율은 0.3487이다.

두 번째는 첫 번째와 동일하게 다변량 정규분포를 가정한 모수적 방법이지만 변수선택법을 이용해 선택된 최적의 변수만을 포함한 방법을 사용한다.

Stepwise Selection Summary										
Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	pop20		0.1223	57.28	<.0001	0.87767577	<.0001	0.12232423	<.0001
2	2	landprice		0.0184	7.69	0.0058	0.86152497	<.0001	0.13847503	<.0001
3	3	pop0		0.0155	6.44	0.0115	0.84816357	<.0001	0.15183643	<.0001
4	4	pop30		0.0150	6.23	0.0130	0.83541039	<.0001	0.16458961	<.0001
5	5	pop60		0.0148	6.11	0.0138	0.82305085	<.0001	0.17694915	<.0001
6	4		pop0	0.0012	0.50	0.4779	0.82407119	<.0001	0.17592881	<.0001
7	5	pop40		0.0080	3.30	0.0702	0.81745127	<.0001	0.18254873	<.0001
8	4		pop30	0.0016	0.66	0.4175	0.81877410	<.0001	0.18122590	<.0001
9	5	subway		0.0070	2.88	0.0905	0.81302162	<.0001	0.18697838	<.0001

Stepwise 변수 선택법에 의해 변수 선택을 진행하였다. pop0 변수와 pop30 변수의 F-value는 각각 0.50, 0.66으로 p-value가 0.5보다 크므로 유의수준 5%에서 유의하지 않은 변수로 판단되었다.

결과적으로 변수 선택을 통해 pop0 변수와 pop30 변수가 제거되었으므로 나머지 pop10, pop20, pop40, pop50, pop60, pop70, subway, landprice 변수를 기준으로 판별함수를 생성해본다.

$X = (X_{pop10}, X_{pop20}, X_{pop40}, \cdots, X_{pop70}, X_{landprice}, X_{subway})$이고 그 외의 수식은 첫 번째 방법과 동일하다.

Number of Observations and Percent Classified into starbucks			
From starbucks	0	1	Total
0	178 84.36	33 15.64	211 100.00
1	93 46.04	109 53.96	202 100.00
Total	271 65.62	142 34.38	413 100.00
Priors	0.5109	0.4891	

Error Count Estimates for starbucks			
	0	1	Total
Rate	0.1564	0.4604	0.3051
Priors	0.5109	0.4891	

재대입법 : 0.3051

Number of Observations and Percent Classified into starbucks			
From starbucks	0	1	Total
0	171 81.04	40 18.96	211 100.00
1	98 48.51	104 51.49	202 100.00
Total	269 65.13	144 34.87	413 100.00
Priors	0.5109	0.4891	

Error Count Estimates for starbucks			
	0	1	Total
Rate	0.1896	0.4851	0.3341
Priors	0.5109	0.4891	

Cross-Validation : 0.3341

재대입법 분류에 의한 오류율은 0.3051이고 교차 타당성에 의한 오분류율은 0.3341이다. 두 가지 오분류율 계산 방법 모두 변수제거를 하지 않았을 때의 오분류율보다 변수 선택에 의해 더 적은 변수를 포함했을 때 더 낮은 오분류율을 나타냈다.

세 번째는 모든 변수를 포함하며 비모수적인 방법인 피셔방법을 사용한 분류이다.

$$X = (X_{pop0}, \cdots, X_{pop70}, X_{landprice}, X_{subway})$$

Number of Observations and Percent Classified into starbucks				
From starbucks	0	1	Other	Total
0	191 90.52	10 4.74	10 4.74	211 100.00
1	114 56.44	69 34.16	19 9.41	202 100.00
Total	305 73.85	79 19.13	29 7.02	413 100.00
Priors	0.5109	0.4891		

Error Count Estimates for starbucks			
	0	1	Total
Rate	0.0948	0.6584	0.3705
Priors	0.5109	0.4891	

재대입법 : 0.3705

Number of Observations and Percent Classified into starbucks			
From starbucks	0	1	Total
0	195 92.42	16 7.58	211 100.00
1	127 62.87	75 37.13	202 100.00
Total	322 77.97	91 22.03	413 100.00
Priors	0.5109	0.4891	

Error Count Estimates for starbucks			
	0	1	Total
Rate	0.0758	0.6287	0.3462
Priors	0.5109	0.4891	

Cross-Validation : 0.3482

재대입법 분류에 의한 오분류율은 0.3705이고 교차 타당성에 의한 오분류율은 0.3482이다. 모수적 방법에 비해서 더 큰 오분류율을 보여준다.

마지막으로 네 번째는 피셔방법을 사용하면서 변수선택법을 기준으로 선택된 최적의 변수들만 포함한다.

$$X = (X_{pop10}, X_{pop20}, X_{pop40}, \cdots, X_{pop70}, X_{landprice}, X_{subway})$$

Number of Observations and Percent Classified into starbucks				
From starbucks	0	1	Other	Total
0	176 83.41	19 9.00	16 7.58	211 100.00
1	110 54.46	78 38.61	14 6.93	202 100.00
Total	286 69.25	97 23.49	30 7.26	413 100.00
Priors	0.5109	0.4891		

Error Count Estimates for starbucks			
	0	1	Total
Rate	0.1659	0.6139	0.3850
Priors	0.5109	0.4891	

재대입법 : 0.3850

Number of Observations and Percent Classified into starbucks			
From starbucks	0	1	Total
0	185 87.68	26 12.32	211 100.00
1	119 58.91	83 41.09	202 100.00
Total	304 73.61	109 26.39	413 100.00
Priors	0.5109	0.4891	

Error Count Estimates for starbucks			
	0	1	Total
Rate	0.1232	0.5891	0.3511
Priors	0.5109	0.4891	

Cross-Validation : 0.3511

재대입법 분류에 의한 오분류율은 0.3850이고 교차 타당성에 의한 오분류율은 0.3511이다. Stepwise 변수 선택법 기준 최적의 변수를 사용한 경우에도 다변량 정규분포를 가정한 모수적 방법이 오분류율 관점에서 더 나은 모형이었다.

결과적으로 변수선택을 통해 pop0과 pop30이 제거된 다변량 정규분포를 가정한 모수적 방법을 최종 분류모형으로 채택하였다.

최종 분류모형에서의 오분류 결과이다.
스타벅스가 입점해있지만 없다고 분류된 동은 다음과 같다.

가락1동	성내3동	명일2동	월계1동	길동	신월3동	보라매동	종암동
가락2동	성북동	목3동	월계2동	길음2동	신정4동	북아현동	중계1동
가양1동	성산2동	목5동	월계3동	남현동	신정6동	불광2동	진관동
가양3동	송중동	묵1동	월곡1동	내곡동	암사3동	사당1동	창1동
갈현1동	송파1동	문정1동	위례동	대조동	약수동	사당3동	창5동
강일동	수서동	미아동	응암1동	도곡2동	양재1동	삼각산동	청파동
개포4동	수유3동	발산1동	이문1동	독산1동	연희동	삼전동	평창동
공릉1동	숭인2동	방배3동	이촌1동	독산3동	오금동	상도3동	홍제1동
광장동	시흥1동	방이1동	일원본동	돈암1동	오류1동	상봉1동	화곡1동
구의2동	신길3동	방화1동	잠실본동	면목7동	왕십리도	상봉2동	화곡4동
군자동	신내1동	번1동	전농1동	명일1동	용신동	서빙고동	흑석동

스타벅스가 없지만 입점해있다고 분류된 동은 다음과 같다.

자양4동	가리봉동	청구동	마천1동	중림동	대림2동	행운동
잠실2동	거여1동	청운효자	상도2동	창신1동	대방동	혜화동
잠실4동	거여2동	풍납2동	서림동	창신3동	동화동	화곡8동
중곡2동	당산1동	행당1동	신길1동	신길7동	신원동	양평1동

스타벅스가 입점해있지만 그렇지 않다고 분류된 경우가 88개, 스타벅스가 없지만 있다고 분류된 동은 28개이다. 입점한 장소에 대한 오분류율이 입점하지 않은 장소에 비해서 상당히 높다.

판별함수는 다음과 같다.

$\overline{X_1} = (\overline{X_{11}}, \overline{X_{12}}, \cdots \overline{X_{1p}})$는 그룹 1(starbucks=0, 스타벅스가 입점하지 않은 동의 그룹)의 평균벡터 행렬이고, $\overline{X_2} = (\overline{X_{21}}, \overline{X_{22}}, \cdots \overline{X_{2p}})$는 그룹 2(starbucks=1, 스타벅스가 입점한 동의 그룹)의 평균벡터 행렬이다.

$$
\begin{aligned}
Q_s(X) &= \ln \frac{f(X|G_1)}{f(X|G_2)} \\
&= \frac{1}{2}\ln\left(\frac{|S_2|}{|S_1|}\right) - \frac{1}{2}(\overline{X_1}' S_1^{-1} \overline{X_1} - \overline{X_2}' S_2^{-1} \overline{X_2}) \\
&+ (\overline{X_1}' S_1^{-1} - \overline{X_2}' S_2^{-1}) X - \frac{1}{2} X' (S_1^{-1} - S_2^{-1}) X
\end{aligned}
$$

142	0	QUAD	pop10	-2.001	-0.161	2.812	-1.030	-0.185	0.491	0.163	-0.019
143	0	QUAD	pop20	-0.161	-3.260	1.745	-0.269	0.216	0.430	0.041	-0.028
144	0	QUAD	pop40	2.812	1.745	-14.574	9.917	0.147	-1.158	-0.376	-0.055
145	0	QUAD	pop50	-1.030	-0.269	9.917	-20.927	12.836	0.026	0.209	0.036
146	0	QUAD	pop60	-0.185	0.216	0.147	12.836	-20.394	7.783	0.116	0.402
147	0	QUAD	pop70	0.491	0.430	-1.158	0.026	7.783	-7.685	-0.441	-0.322
148	0	QUAD	subway	0.163	0.041	-0.376	0.209	0.116	-0.441	-0.558	-0.014
149	0	QUAD	landprice	-0.019	-0.028	-0.055	0.036	0.402	-0.322	-0.014	-0.535
150	0	QUAD	_LINEAR_	0.645	-1.226	-2.611	0.447	1.583	0.043	0.046	-0.164
151	0	QUAD	_CONST_	5.162	5.162	5.162	5.162	5.162	5.162	5.162	5.162

$$f(X|G_1) = 5.162 + (0.645,\ -1.226,\ -2.611, 0.447, 1.583, 0.043, 0.046,\ -0.164)\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_8 \end{pmatrix}$$

$$+ (X_1 X_2 \cdots X_8)\begin{pmatrix} -2.001 & -0.161 & 2.812 & -1.030 & -0.185 & 0.491 & 0.163 & -0.019 \\ -0.161 & -3.260 & 1.745 & -0.269 & 0.216 & 0.430 & 0.041 & -0.028 \\ 2.812 & 1.745 & -14.574 & 9.917 & 0.147 & -1.158 & -0.376 & -0.055 \\ -1.030 & -0.269 & 9.917 & -20.927 & 12.836 & 0.026 & 0.209 & 0.036 \\ -0.185 & 0.216 & 0.147 & 12.836 & -20.394 & 7.783 & 0.116 & 0.402 \\ 0.491 & 0.430 & -1.158 & 0.026 & 7.783 & -7.685 & -0.441 & -0.322 \\ 0.163 & 0.041 & -0.376 & 0.209 & 0.116 & -0.441 & -0.558 & -0.014 \\ -0.19 & -0.028 & -0.055 & 0.036 & 0.402 & -0.322 & 0.046 & -0.535 \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_8 \end{pmatrix}$$

152	1	QUAD	pop10	-0.855	0.214	-0.078	0.010	0.767	-0.233	0.026	-0.112
153	1	QUAD	pop20	0.214	-0.823	0.501	-0.021	-0.205	0.186	-0.119	0.012
154	1	QUAD	pop40	-0.078	0.501	-3.757	5.862	-2.555	-0.116	0.053	-0.081
155	1	QUAD	pop50	0.010	-0.021	5.862	-15.754	10.228	-0.195	-0.308	0.207
156	1	QUAD	pop60	0.767	-0.205	-2.555	10.228	-14.618	6.554	0.291	0.426
157	1	QUAD	pop70	-0.233	0.186	-0.116	-0.195	6.554	-6.755	-0.321	-0.488
158	1	QUAD	subway	0.026	-0.119	0.053	-0.308	0.291	-0.321	-0.900	-0.028
159	1	QUAD	landprice	-0.112	0.012	-0.081	0.207	0.426	-0.488	-0.028	-0.559
160	1	QUAD	_LINEAR_	-0.039	0.213	0.270	-0.051	-0.225	-0.017	-0.148	0.149
161	1	QUAD	_CONST_	3.402	3.402	3.402	3.402	3.402	3.402	3.402	3.402

$$f(X|G_1) = 3.402 + (-0.039, 0.213, 0.270,\ -0.051,\ -0.225,\ -0.017,\ -0.148, 0.149)\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_8 \end{pmatrix}$$

$$+ (X_1 X_2 \cdots X_8)\begin{pmatrix} -0.855 & 0.214 & -0.078 & 0.010 & 0.767 & -0.233 & 0.026 & -0.112 \\ 0.214 & -0.823 & 0.501 & -0.021 & -0.205 & 0.186 & -0.119 & 0.012 \\ -0.078 & 0.501 & -3.757 & 5.862 & -2.555 & -0.116 & 0.053 & -0.081 \\ 0.010 & -0.021 & 5.862 & -15.754 & 10.288 & -0.195 & -0.308 & 0.207 \\ 0.767 & -0.205 & -2.555 & 10.228 & -14.618 & 6.554 & 0.291 & 0.426 \\ -0.233 & 0.186 & -0.116 & -0.195 & 6.554 & -6.755 & -0.321 & -0.488 \\ 0.026 & -0.119 & 0.053 & -0.308 & 0.291 & -0.321 & -0.900 & -0.028 \\ -0.112 & 0.012 & -0.081 & 0.207 & 0.426 & -0.488 & -0.028 & -0.559 \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_8 \end{pmatrix}$$

$$Q_s(X) = \ln\frac{f(X|G_1)}{f(X|G_2)} > \ln(\frac{p_1}{p_2}) \simeq 0\ (p_1 \simeq p_2)$$이면 X를 G_1으로 분류하고 그렇지 않으면 G_2으로 분류한다.

다음 세 동은 지가 변수에 결측이 존재하여 분석에서 제외된 동들이다. 결측값을 인접한 동의 지가 변수로 대체하여 최종 분류모형에 대입시켜서 실제의 스타벅스

유무와 비교한다.

(표준화)	pop10	pop20	pop40	pop50	pop60	pop70	subway	landprice
낙성대동	-0.441	0.502	-0.759	-0.839	-0.776	-0.774	-0.242	0.295
응봉동	0.443	-0.377	0.413	-0.015	0.095	0.283	-0.517	-0.783
이촌2동	-0.553	-0.686	-0.239	-0.550	-0.403	0.386	0.324	-0.513

Posterior Probability of Membership in starbucks

dong	From starbucks	Classified into starbucks		0	1
낙성대동	.	1	*	0.1541	0.8459
응봉동	.	0	*	0.5449	0.4551
이촌2동	.	1	*	0.2922	0.7078

낙성대동은 분류 결과 그룹 0일 확률이 0.1541, 그룹 1일 확률이 0.8459로 스타벅스가 입점할 만한 동으로 분류되었다. 응봉동은 그룹 0일 확률이 0.5449, 그룹 1일 확률이 0.4551로 스타벅스가 입점하지 않을 동으로 분류되었다. 이촌2동은 그룹 0일 확률이 0.2922, 그룹 1일 확률이 0.7078로 스타벅스가 입점할 동으로 분류되었다.

세 동 모두 실제로 스타벅스가 입점해있는 동으로 낙성대동과 이촌2동은 잘 분류되었지만 응봉동은 오분류가 발생하였다. 응봉동의 경우, 각 그룹에 속할 확률이 큰 차이가 나지 않으므로 완벽히 실패한 분류로 볼 수는 없다.

2.3 분석의 타당성 설명

Error Count Estimates for starbucks

	0	1	Total
Rate	0.1564	0.4604	0.3051
Priors	0.5109	0.4891	

Error Count Estimates for starbucks

	0	1	Total
Rate	0.1896	0.4851	0.3341
Priors	0.5109	0.4891	

재대입법 : 0.3051 Cross-Validation : 0.3341

비교한 네 개의 모델 중 오분류율을 기준으로 가장 작은 모형을 사용하였다.

Multivariate Statistics and Exact F Statistics					
S=1 M=3 N=201					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.80984172	11.86	8	404	<.0001
Pillai's Trace	0.19015828	11.86	8	404	<.0001
Hotelling-Lawley Trace	0.23480919	11.86	8	404	<.0001
Roy's Greatest Root	0.23480919	11.86	8	404	<.0001

판별함수의 유의성 검정을 시행하였다. 귀무가설은 두 그룹의 평균벡터가 동일하다는 것이다. 그 결과 Wilks' Lambda 통계량 값이 약 0.8098로 p-value는 <0.0001이다. 따라서 귀무가설을 기각할 수 있게 되고 판별함수는 두 그룹을 유의하게 분리한다고 할 수 있다.

3. 결론

3.1 분석 결과 요약

총 네 가지 모형을 비교하여 최종 분류모형으로 변수 선택을 통해 pop0과 pop30이 제거된 다변량 정규분포를 가정한 모수적 방법이 최종 분류모형으로 채택되었다. 두 그룹의 공분산 행렬이 다르기 때문에 그에 맞는 이차형식의 판별함수를 얻었으며 다음과 같으며 오분류율은 재대입법 기준으로 약 0.3이다.

$$f(X|G_1) = 5.162 + (0.645, -1.226, -2.611, 0.447, 1.583, 0.043, 0.046, -0.164)\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_8 \end{pmatrix}$$

$$+ (X_1 X_2 \cdots X_8)\begin{pmatrix} -2.001 & -0.161 & 2.812 & -1.030 & -0.185 & 0.491 & 0.163 & -0.019 \\ -0.161 & -3.260 & 1.745 & -0.269 & 0.216 & 0.430 & 0.041 & -0.028 \\ 2.812 & 1.745 & -14.574 & 9.917 & 0.147 & -1.158 & -0.376 & -0.055 \\ -1.030 & -0.269 & 9.917 & -20.927 & 12.836 & 0.026 & 0.209 & 0.036 \\ -0.185 & 0.216 & 0.147 & 12.836 & -20.394 & 7.783 & 0.116 & 0.402 \\ 0.491 & 0.430 & -1.158 & 0.026 & 7.783 & -7.685 & -0.441 & -0.322 \\ 0.163 & 0.041 & -0.376 & 0.209 & 0.116 & -0.441 & -0.558 & -0.014 \\ -0.19 & -0.028 & -0.055 & 0.036 & 0.402 & -0.322 & 0.046 & -0.535 \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_8 \end{pmatrix}$$

$$f(X|G_1) = 3.402 + (-0.039, 0.213, 0.270, -0.051, -0.225, -0.017, -0.148, 0.149)\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_8 \end{pmatrix}$$

$$+ (X_1\, X_2 \cdots X_8)\begin{pmatrix} -0.855 & 0.214 & -0.078 & 0.010 & 0.767 & -0.233 & 0.026 & -0.112 \\ 0.214 & -0.823 & 0.501 & -0.021 & -0.205 & 0.186 & -0.119 & 0.012 \\ -0.078 & 0.501 & -3.757 & 5.862 & -2.555 & -0.116 & 0.053 & -0.081 \\ 0.010 & -0.021 & 5.862 & -15.754 & 10.288 & -0.195 & -0.308 & 0.207 \\ 0.767 & -0.205 & -2.555 & 10.228 & -14.618 & 6.554 & 0.291 & 0.426 \\ -0.233 & 0.186 & -0.116 & -0.195 & 6.554 & -6.755 & -0.321 & -0.488 \\ 0.026 & -0.119 & 0.053 & -0.308 & 0.291 & -0.321 & -0.900 & -0.028 \\ -0.112 & 0.012 & -0.081 & 0.207 & 0.426 & -0.488 & -0.028 & -0.559 \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_8 \end{pmatrix}$$

3.2 분석의 장점 및 한계점 설명

본 연구의 모형을 이용하면 특정 동의 몇 가지 변수만을 간단히 판별함수에 대입하여 스타벅스의 입점 여부를 추정해볼 수 있다. 또한, 표준화 변수를 이용하였기 때문에 변수들끼리 단위의 차이가 있더라도 변수들을 동등하게 분류에 반영할 수 있다.

한계점으로는 우선 한 동에 여러 매장이 입점한 경우가 존재하는데 본 연구의 모형은 이런 경우를 고려할 수 없다. 스타벅스의 정책과 관련된 문제인데 모든 매장이 직영점으로 운영되기 때문에 다른 매장과의 거리는 크게 중요한 요인이 아니다. 거리가 가깝더라도, 심지어 같은 동에 있더라도 예측 수요량이 높다면 매장을 입점시킨다.

또한, 스타벅스의 입점에는 사용한 변수 외에도 더 다양한 요인이 존재할 것이다. 더 많은 타당한 변수를 사용한다면 모형의 성능을 개선할 수 있을 것이다. 예를 들면 해당 동의 주거지역인지 상업지역인지 같은 지역의 특성을 반영하는 용도지역을 변수로 사용해볼 만한 가치가 있다고 생각된다.

개별 상권 및 상가의 특성을 고려하지 못한 점도 한계점이다. 물론 지하철과 각 상가와의 평균 거리를 고려했지만, 이는 사람이 많이 모이며 약속장소로 널리 이용되는 지하철역의 특성을 반영하기 위한 변수이며 상가의 특성과는 거리가 멀다.

3.3 추가 연구사항 제안

본 연구는 서울시의 행정동을 기준으로 분석을 진행하였다. 각 상가들을 기준으로 분석을 진행한다면 더 정밀한 결과를 얻을 수 있을 것이다. 또, 스타벅스는 서울시 외에도 한국에 많은 매장을 내고 있다. 전국을 기준으로 한 분석도 제안한다.

4. 참고문헌

1. 서울시립대학교 통계학과 김규성 교수님 다변량 통계학 강의자료
2. 서울시 프랜차이즈 커피점 입지의 영향요인 -스타벅스 커피전문점을 중심으로-, 정승영, 최인섭
3. 위치정보를 활용한 커피 전문점의 입점 분석, 이동엽, 윤영태
4. SAS HELP CENTER, PROC DISCRIM Statement https://documentation.sas.com/?cdcId=pgmsascdc&cdcVersion=9.4_3.3&docsetId=statug&docsetTarget=statug_discrim_syntax01.htm&locale=en
4. 서울시 행정동단위 월별 KT 유동인구, 서울특별시 빅데이터 캠퍼스, https://bigdata.seoul.go.kr/mobile/data/selectSampleData.do?r_id=M310&sample_data_seq=73&tab_type=&sch_cate=10&file_id=&sch_text=&sch_order=U&currentPage=1
5. 지가 변동률 데이터, 한국 감정원 부동산 통계 정보 시스템 (https://www.r-one.co.kr/)
6. 지하철역 정보 데이터, 서울 열린 데이터 광장 https://data.seoul.go.kr/dataList/32/literacyView.do
7. 상가정보 데이터(소상공인시장진흥공단_상가(상권)정보_서울_20200930), 공공데이터 포털(https://www.data.go.kr/data/15059995/fileData.do)
8. SAS를 이용한 다변량 통계 분석