Data Analytics
Individual Assignment 1

202011431 산업공학과 차승현

다중 회귀분석기법을 이용한 위생학 분야 특허의 피인용 수 예측

건국대학교

목차

I. 서론

기업과 국가 경제의 성장에 있어 기술의 중요성이 증가하고, 시장에서의 경쟁력 제고를 위해 지식재산 확보 경쟁이 심화되고 있다. 이에 따라, 기업의 경영 전략과 국가 경제 정책에서 지식재산 정보의 활용이 증가하고 있다. 지식재산 정보는 출원·등록된 지식재산의 기술적 세부 정보와, 권리자 및 권리 내용 등 다양한 정보를 포함하고 있어 새로운 지식재산을 창출하는 과정에서 신규성과 진보성을 확보·검토하는 선행기술 조사 및 동향분석에 활용되고, 기업의 사업 전략 수립시 관련 기술의 탐색이나 지식재산 확보를 위한 정보 분석에 활용되었다. 또한 창업, 신사업 탐색 등 지식재산 기반의 경영 활동에서 활용이 증가하고 있으며, 지식재산 정보는 국내에서만 연간 출원 건수가 48만 건에 이르고 누적 출원 수는 천만 건을 돌파하였다 (특허청 2020).

그러나, 한국 데이터베이스진흥원(2010)의 데이터 품질관리 성숙수준 조사 보고서에 따르면 데이터의 품질을 효율적으로 통제하기 위한 품질관리 방안이 마련되어 있지 않거나, 데이터 품질에 대한 인식 수준이 미흡한 수준으로 나타났다.

특허의 품질에서 특허의 피인용 수는 특허의 가치와 긍정적이고 유의미한 관계가 있는 것으로 나타났으며(Carpenters & Woolf, 1984), 피인용 분석은 소송, 특허 양도 가격의 책정 등의 목적으로 특허를 평가할 때 점점 더 일반적으로 사용되고 있음을 확인할 수 있다(Nathan Falk et al., 2016).

자주 인용되는 특허는 보다 높은 기술적, 경제적인 가치를 창출하며(Dutta & Weiss, 1997) 발명의 market value 추정치가 높을수록 특허의 인용 횟수가 더 많아짐을 확인한 연구 사례가 존재한다(D Harhoff et al., 1999). 또한, 특허당 피인용이 증가할수록 market value가 대략 3% 증가한다고 알려졌다 (Hall & Trajtenberg, 2005).

특허의 품질에 영향을 끼치는 수많은 요소들 중, 대표적으로 특허의 피 인

용수와 해당 특허의 품질은 비례관계에 있다고 판단하였기에 본 보고서에서는 특허의 인용에 영향을 미치는 몇 가지 변수들을 토대로 피 인용횟수를 예측하고자 한다.

II. 분석 기법

회귀분석은 종속변수(Dependent Variable)와 독립변수(Independent Variable)간의 상관관계를 검증하여 독립변수가 종속변수에 어떠한 영향력을 미치는지 파악하거나, 독립변수의 변화에 따라 종속변수의 변화를 예측하기 위하여 사용되는 통계학적 분석방법이다. 회귀분석은 독립변수의 개수에 따라 독립변수가 둘 이상인 경우는 다중회귀분석, 하나인 경우는 단순회귀분석이라 한다. 회귀분석이 사용되는 이유는 결과(종속변수)의 일부 원인(독립변수)을 한 번에 분석이 가능하기 때문이다. 또한 회귀분석에서는 종속변수에 대한 각각의 독립변수들이 어떠한 영향을 미치는지 개별적으로 분석이 가능하기 때문에 특정 변수를 통제할 시 다른 독립변수가 종속 변수의 변화에 어떠한 상호관련성이 있는지 쉽게 판단이 가능하다. 하지만, 독립변수간의 상호연관성을 배제하고, 단방향의 관계만을 취급하는 특징을 갖고 있다. 또한 측정 오차를 허용하지 않는 특징을 가지고 있다. 따라서 독립변수내의 관련성 문제 및 다중공선성 문제를 극복 가능한 간단한 인과모형을 대상으로 할 시, 종속변수에 대한 독립변수간의 상호영향력의 크기를 비교 가능한 뛰어난 통계기법 중 하나이다. 다중회귀분석에서 사용된 수식은 아래와 같고, Y_i는 종속변수인 특허의 피 인용수를 의미하며, 각 X_{ni}는 독립변수들을 의미한다.

$$Y_i = B_1 + B_2X_{2i} + B_3X_{3i} + \ldots + B_{10\sum}X_{10i} + u_i(i = 1, 2, \ldots, n)$$

III. 데이터 정제 방안 및 분석 대상

기존 선행적으로 특허의 품질에 관련된 연구와 일반적으로 특허의 가치를평가하기 위해 사용되는 변수는 크게 다음과 같이 요약할 수 있다.

번호	변수명	변수 설명
1	발명자 수	특허 발명에 참여한 발명자의 수
2	승인 기간	출원 신청부터 등록 승인까지의 기간
3	특허 범위	* 특허의 IPC 수 * IPC는 기술의 범위를 나타냄. IPC 수가 많을수록 다양한 기술이 사용되었음을 의미 *고품질 특허와 IPC code 수 사이의 양의 관계를 발견(JL Wu et al., 2016)
4	청구항 수	* 특허의 청구항 수 * 청구항 수가 많다는 것은 보다 많은 발명, 즉 넓은 기술적 범위에 대한 보호가 선언된 것으로 간주할 수 있음. * 청구항 수가 기술혁신 활동에 대한 보다 정확한 정보를 제공(X.Tong & J.D.Frame, 1994)
5	청구항 길이	청구항 문자열의 길이
6	설명서 길이	발명 설명서 문자열 길이
7	이미지 수	특허에 포함된 이미지의 수
8	인용 특허 수	특허가 인용한 참고 문헌의 수
9	비특허 인용수	*특허가 인용한 비특허 문헌(논문) * 특허의 기술이 과학 연구성과와 얼마나 밀접한 관계를 가지는지 나타냄
10	특허 패밀리	* 특허가 출원된 국가의 수 * 특허의 지역적 보호 범위를 나타내며, 간접적으

		로는 해당 특허가 가지는 기술적 중요성과 혁신성과로서의 가치에 대한 정보를 제공(Harhoff et al., 2003)
11	특허권자의 피인용 수	* 특허권자가 출원한 모든 특허들의 피인용 수 * 특허권자의 기술적 수준이 높을수록 해당 특허의 가치가 높을 것
12	특허권자의 평균 피인용 수	* 특허권자가 출원한 모든 특허들의 피인용 수의 평균 * 특허를 많이 등록하지는 않았지만 소수의 고품질 특허를 등록한 특허권자를 구분할 수 있을 것
13	인용 특허의 피인용 수	* 특허가 인용한 참고 문헌들의 총 피인용 수 * 참고 문헌들의 가치가 높을수록 해당 특허의 가치가 높을 것으로 예상
14	인용 특허의 평균 피인용 수	* 특허가 인용한 참고 문헌들의 피인용 수의 평균
15	발명자의 피인용 수	* 특허 발명에 참여한 발명자가 등록한 특허들의 총 피인용 수
16	발명자의 평균 피인용 수	* 특허 발명에 참여한 발명자가 등록한 특허들의 총 피인용 수

이외에도, 타 연구에서 사용된 데이터는 출원인 국적, 출원인 유형, 출원국가 수, 참고문헌 평균 피인용횟수, 서지결합도 등 여러 변수가 있었지만 보유한 파일데이터 내에서는 위와 같은 변수를 사용할 수 없다고 판단하였다.

본 과제에서 주어진 ipc, abstract, inventor, assignee, us_patent, citation에서 이용할 수 있는 독립변수는 다음과 같다.

파일	가용 독립변수
inventor.txt	1. 발명자 수

us_patent.txt	2. 승인 기간 3. 특허 범위 도면 수

1. 발명자 수

Inventor.txt에서 '특허등록번호'를 기준으로 동일한 특허등록번호가 n개의 row가 존재한다면, 발명자 수를 n명으로 추출해낼 수 있다.

2. 승인 기간

us_patent.txt에서 '특허등록일자'와 '특허출원일자'의 datatype을 datetime으로 변환 후 빼면 승인 기간을 추출해낼 수 있다.

3. 특허 범위

Us_patent.txt에서 '보유IPC수'와 '보유CPC수'를 추출할 수 있다. 해당 파일에서 '보유USPC수' column은 모든 값이 0이므로 제외한다.

본 보고는 '특허의 피인용 수'를 예측하는 기법을 산정하기 위한 보고로, 인용 수와 관련된 변수는 독립변수로 선정할 수 없다고 판단했다. 추가로, 가용이 가능할 것으로 간주되는 도면 수를 독립변수로 선정했다.

변수구분	변수명
종속변수	특허의 피인용 수
독립변수	발명자 수
	승인 기간
	보유 IPC 수
	보유 CPC 수
	도면 수

여기서, 분석하고자 하는 데이터는 위생학, 의학 또는 수의학 분야의 특허이므로 ipc.csv 파일에서 '보유IPC전체코드' column의 앞 3글자가 A61로 시

작되는 특허 데이터로 한정하였다.

IV. 분석 과정

	특허등록번호	보유IPC전체코드
172	9532550	A61D-009/00
176	9532552	A61D-001/02
256	9532570	A61K-031/513
263	9532571	A61K-008/365
314	9532585	A61K-009/48
...	...	...
1563609	9854101	A61F-011/06
1564061	9854203	A61B-005/01
1565225	9854356	A61F-011/06
1565287	9854370	A61N-001/00
1566914	9854656	A61B-006/00

37580 rows × 2 columns

1) ipc.csv에서 '보유IPC전체코드' Column을 기준으로, 첫 3글자가 A61로 시작하는(위생학 분야의 IPC 분류체계 클래스) 것들을 뽑아내고 하나의 특허등록번호에 여러 개의 보유 IPC전체코드를 가질 수 있으므로, 특허등록번호를 unique화하여 1:1 관계를 생성한다. 이후 ipc_2라는 변수에 저장한다.

	특허등록번호	특허등록일자	특허출원일자	보유IPC수	보유CPC수	도면수	특허의후방인용수
0	9532550	20170103	20130903	2	1	16	3
1	9532552	20170103	20130402	2	2	5	26
2	9532570	20170103	20141230	9	7	0	31
3	9532571	20170103	20150710	16	20	0	37
4	9532585	20170103	20020116	5	5	0	38
...	...	...	...	...	...	...	...
37575	9854101	20171226	20160210	11	15	18	38
37576	9854203	20171226	20160506	11	36	20	2
37577	9854356	20171226	20150507	5	9	7	19
37578	9854370	20171226	20170228	7	12	7	20
37579	9854656	20171226	20140904	6	6	5	18

2) us_patent.txt를 dataframe 타입으로 불러온 후 ipc_2와 '특허등록번호'를 기준으로 inner join하고 사용하지 않을 변수인 '특허출원번호', '특허제목', '보유USPC수' column을 제거한다. 이를 df_ver2라는 변수에 저장한다.

	특허등록번호	발명자수
0	9762553	53
1	9613190	53
2	9593129	42
3	9727622	36
4	9734217	36
...	...	...
314785	9754438	1
314786	9754436	1
314787	9596946	1
314788	9596947	1
314789	9765719	1

3) inventor.txt를 dataframe 타입으로 불러온 후 value_counts() 모듈을 활용하여 고유한 특허등록번호에 기여한 발명자의 수를 추출하고, 이를 특허등록번호를 기준으로 df_ver2와 inner join한 후 df_ver3이라는 변수에 저장한다.

	특허등록번호	특허등록일자	특허출원일자	보유IPC수	보유CPC수	도면수	특허의후방인용수	발명자수
0	9532550	20170103	20130903	2	1	16	3	1
1	9532552	20170103	20130402	2	2	5	26	1
2	9532570	20170103	20141230	9	7	0	31	3
3	9532571	20170103	20150710	16	20	0	37	2
4	9532585	20170103	20020116	5	5	0	38	2
...	...	...	...	...	...	...	...	...
37575	9854101	20171226	20160210	11	15	18	38	4
37576	9854203	20171226	20160506	11	36	20	2	3
37577	9854356	20171226	20150507	5	9	7	19	1
37578	9854370	20171226	20170228	7	12	7	20	1
37579	9854656	20171226	20140904	6	6	5	18	2

[df_ver3]

	특허등록번호	특허등록일자	특허출원일자	보유IPC수	보유CPC수	도면수	특허의후방인용수	발명자수	특허승인소요기간
0	9532550	20170103	20130903	2	1	16	3	1	1218
1	9532552	20170103	20130402	2	2	5	26	1	1372
2	9532570	20170103	20141230	9	7	0	31	3	735
3	9532571	20170103	20150710	16	20	0	37	2	543
4	9532585	20170103	20020116	5	5	0	38	2	5466
...	...	...	...	...	...	...	...	...	...
37575	9854101	20171226	20160210	11	15	18	38	4	685
37576	9854203	20171226	20160506	11	36	20	2	3	599
37577	9854356	20171226	20150507	5	9	7	19	1	964
37578	9854370	20171226	20170228	7	12	7	20	1	301
37579	9854656	20171226	20140904	6	6	5	18	2	1209

4) 특허등록일자와 특허출원일자 column의 date type이 int형이므로, datetime형으로 변환한 후, 특허등록일자 - 특허출원일자 = 특허승인소요기간으로 설정하여 소요기간을 산출하고, 특허승인소요기간을 다시 수치형으로 변환해 df_ver5라는 변수에 저장한다.

	특허등록번호	보유IPC수	보유CPC수	도면수	발명자수	특허승인소요기간	특허의후방인용수
0	9532550	2	1	16	1	1218	3
1	9532552	2	2	5	1	1372	26
2	9532570	9	7	0	3	735	31
3	9532571	16	20	0	2	543	37
4	9532585	5	5	0	2	5466	38
...	...	...	...	...	...	...	...
37575	9854101	11	15	18	4	685	38
37576	9854203	11	36	20	3	599	2
37577	9854356	5	9	7	1	964	19
37578	9854370	7	12	7	1	301	20
37579	9854656	6	6	5	2	1209	18

5) 특허등록일자와 특허출원일자는 특허승인소요기간을 위한 변수이므로 제거한 후, 종속변수에 해당하는 특허의후방인용수 column을 제일 뒤쪽으로

setting한 후 df_ver7라는 변수에 저장하였고, 해당 데이터를 사용하여 다중회귀분석을 실시한다.

Dep. Variable:	특허의후방인용수	**R-squared:**	0.119
Model:	OLS	**Adj. R-squared:**	0.119
Method:	Least Squares	**F-statistic:**	1012.
Date:	Thu, 31 Mar 2022	**Prob (F-statistic):**	0.00
Time:	02:45:50	**Log-Likelihood:**	-2.6600e+05
No. Observations:	37580	**AIC:**	5.320e+05
Df Residuals:	37574	**BIC:**	5.321e+05
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.8761	4.061	0.216	0.829	-7.084	8.837
보유IPC수	-5.9196	0.459	-12.888	0.000	-6.820	-5.019
보유CPC수	5.5753	0.282	19.742	0.000	5.022	6.129
도면수	3.0910	0.053	58.114	0.000	2.987	3.195
발명자수	4.3806	0.633	6.917	0.000	3.139	5.622
특허승인소요기간	0.0033	0.002	1.649	0.099	-0.001	0.007

Omnibus:	55591.447	**Durbin-Watson:**	1.577
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	23495385.405
Skew:	9.122	**Prob(JB):**	0.00
Kurtosis:	124.129	**Cond. No.**	3.92e+03

모든 변수를 적합시킨 다중회귀분석의 결과

6) R-sqaured, Adj.R-squared의 값이 0.119로 설명력이 매우 낮다고 해석할 수 있다. 추가로, p-value가 constant, 특허승인소요기간 column에서 0.05보다 큰 값을 보이고 있어 통계적으로 유의하다고 할 수 없다. 따라서, 전진선택법(Forward Selection), 후진제거법(Backward Elimination), 단계적 선택법(Stepwise Selection)을 사용하여 변수를 선택하고자 한다.

7) 전진선택법 적용

전진 선택법은 기존 모형에 가장 설명력이 좋은 변수를 하나씩 추가하는 방법이다. 전진 선택법에서는 변수의 추가 여부를 결정하는 유의수준을 결정하는데, 가장 설명력이 좋은 변수이더라도 해당 유의수준을 만족하지 못할 경우, 선택되지 못하고 전진 선택 알고리즘은 끝나게 된다.

S를 기존 모형에 포함된 변수들의 집합, $\tilde{S}$를 모형에 포함되지 않은 변수들의 집합, 그리고 α를 유의수준이라고 하자.

7-1) 아직 모형에 적합시키지 않은 변수 $X_k \in \tilde{S}$를 기존 모형에 추가하여 적합한다. 기존 모형에 추가하여 적합한다는 말은, 기존 모형에 있던 변수와 추가된 변수 X_k 를 이용하여 선형 모형을 적합한다는 의미이다.

7-2) 변수 X_k에 대한 회귀계수 b_k를 구하고, b_k에 대한 t통계량을 계산한다. 이후 t통계량에 대응하는 p-value를 구한다. 이 작업을 $\tilde{X}$에 있는 모든 변수에 대하여 수행한다.

이때 t 통계량은 다음과 같다.

$$t = \left(\frac{SSR(X_k \cup S) - SSR(S)}{1} \div \frac{SSE(X_k \cup S)}{n - card(S) - 1} \right)^{\frac{1}{2}}$$

여기서 SSR(A), SSE(A)는 각각 집합 A에 포함된 모든 변수를 포함한 선형 모형의 회귀 제곱합과 잔차의 제곱합이며, card(A)는 집합 A의 변수 개수이다. 또한 t통계량은 자유도가 n – card(S) – 1인 t분포를 따르므로 p-value는 다음과 같이 구한다.

$$p\text{-value} = P(T > t)$$

7-3) 이때 최소 p-value 값과 미리 정해둔 유의수준 α와 비교한다. 만약 최소 p-value < α이면, 최소 p-value 에 해당하는 변수를 S에 포함시키고 7-1, 7-2 단계를 수행한다. 그렇지 않은 경우에는 알고리즘을 종료한다.

전진선택법은 변수가 많은 상황에서도 사용이 가능하다는 장점이 있지만, 한번 선택된 변수는 계속 모형에 존재하고 일치성(샘플수가 많아질수록 실제 모형에 수렴하는 성질)이 만족되지 않는다는 단점이 있다.

전진선택법을 사용하여 최종적으로 선택된 독립변수는 '도면수', '보유 CPC 수', '보유 IPC 수', '발명자수' 변수이다.

```
selected_variables
```

```
['도면수', '보유CPC수', '보유IPC수', '발명자수']
```

8) 후진제거법 적용

후진제거법은 전진선택법과는 반대로, 모든 변수가 포함된 모형에서 설명력이 가장 적은 변수를 제거해가는 기법이다.

8-1) 현재 모형에 포함된 변수를 이용하여 선형 모형을 적합한다.

8-2) 추정된 절편항을 제외한 회귀 계수에 대하여 가장 큰 p-value 의 값을 구한다.

8-3) 만약 p-value 의 최대값이 사전에 정의한 유의수준 α보다 크다면 최대 p-value 에 대응하는 변수를 기존 모형에서 제외한다. 그렇지 않다면 후진 제거법 알고리즘을 종료한다.

후진제거법도 마찬가지로, 변수가 많은 데이터에 적용이 가능하지만 한번 제외된 변수는 다시 모형에 포함될 수 없고 일치성을 만족하지 않는다는 단점이 존재한다.

후진제거법을 사용하여 최종적으로 선택된 독립변수는 '보유 IPC 수', '보유 CPC 수', '도면수', '발명자수' 변수이다.

```
selected_variables
```

```
['보유IPC수', '보유CPC수', '도면수', '발명자수']
```

9) 단계적 선택법 적용

단계적 선택법은 전진 선택법에서 후진 제거법을 추가한 방법이다. 9-1)과 9-2)는 전진 선택법의 7-1)과 7-2)와 동일하다.

9-3) 최소 p-value 의 값과 미리 정해둔 유의수준 α와 비교한다. 만약 최소 p-value < α이면 최소 p-value 에 해당하는 변수를 S 에 포함시키고 9-4)단계로 넘어간다. 그렇지 않다면 알고리즘을 종료한다.

9-4) 추가된 변수를 포함하여 현재 S 에 있는 모든 변수를 이용해 선형 모형을 적합한다. 그리고 절편항을 제외한 추정된 회귀 변수에 대하여 가장 큰 p-value 값을 구한다.

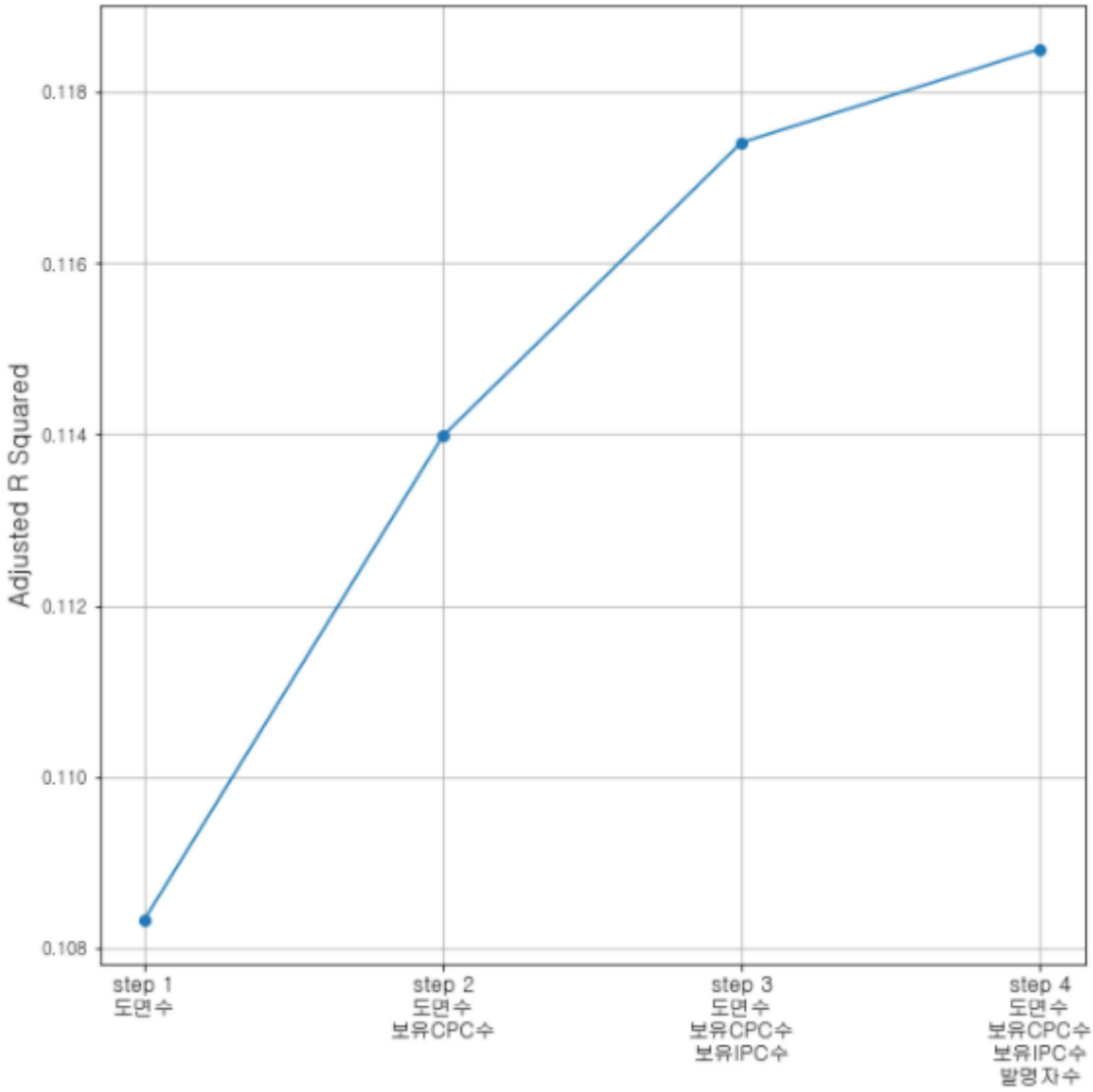
9-5) 최대 p-value 값이 사전에 정의된 유의수준보다 크거나 같다면, 해당 변수를 제외하고 1 단계로 넘어간다. 그렇지 않다면 제외하는 변수 없이 바로 9-1)로 돌아간다.

단계적 선택법은 한 번 들어간 변수는 계속 포함된다는 전진 선택법의 단점을 보완했다는 장점이 있다. 하지만 변수가 많아지면 계산량이 늘어나고, 일치성을 만족시키지 않는다면 단점이 있다.

단계적 선택법을 사용하여 최종적으로 선택된 독립변수는 '도면수', '보유 CPC 수', '보유 IPC 수', '발명자수' 변수이다.

```
selected_variables
```

```
['도면수', '보유CPC수', '보유IPC수', '발명자수']
```



적합의 정도가 점점 좋아지는 것을 확인할 수 있다. 전진선택법, 후진제거법, 단계적 선택법 모든 변수 선택법의 결과가 '특허승인소요기간'을 제외한 column 선택이 효율적이라는 결과가 도출되었으므로, 이를 제외한 4가지 변수를 사용하여 다시 다중회귀분석을 시도한다.

10) 다중회귀분석 2번째 시도

Dep. Variable:	특허의후방인용수	R-squared:	0.119
Model:	OLS	Adj. R-squared:	0.119
Method:	Least Squares	F-statistic:	1264.
Date:	Thu, 31 Mar 2022	Prob (F-statistic):	0.00
Time:	02:22:29	Log-Likelihood:	-2.6600e+05
No. Observations:	37580	AIC:	5.320e+05
Df Residuals:	37575	BIC:	5.321e+05
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.2616	3.070	1.714	0.087	-0.756	11.279
보유IPC수	-5.9541	0.459	-12.976	0.000	-6.853	-5.055
보유CPC수	5.5640	0.282	19.707	0.000	5.011	6.117
도면수	3.0898	0.053	58.096	0.000	2.986	3.194
발명자수	4.3712	0.633	6.902	0.000	3.130	5.612

Omnibus:	55595.095	Durbin-Watson:	1.577
Prob(Omnibus):	0.000	Jarque-Bera (JB):	23500518.602
Skew:	9.123	Prob(JB):	0.00
Kurtosis:	124.142	Cond. No.	75.5

특허승인소요기간 변수를 제외하고 다중회귀분석을 시도하였음에도 불구하고, 수정된 결정계수와 결정계수 모두 0.119로 설명력이 낮다는 결론에 도달하였다.

11) 다중공선성(VIF) 확인

다중공선성이 존재하게되면 회귀계수의 추정량의 분산이 커지게 된다. 회귀계수 추정량의 분산이 커지면 추정한 회귀계수의 정확성이 떨어지게 되고, 이에 따라 회귀계수 추정값을 신뢰할 수 없게 된다.

다중공선성은 분산팽창인자(Variance Inflation Factor)의 값을 보고 확인할 수 있다. 설명변수 x_k의 분산팽창인자 VIF_k는 다음과 같이 정의한다.

$$VIF_k = (1 - R_k^2)^{-1}$$

VIF_k의 값이 10보다 크다면, 다중공선성이 존재한다고 판단하게 된다.

	VIF Factor	features
0	5.468061	보유IPC수
1	5.492260	보유CPC수
2	1.531062	도면수
3	2.096624	발명자수

독립변수의 분산팽창인자

해당 데이터는 다중공선성을 위배한다고 볼 수 없었다. 그러나, 상관관계를 확인해 보았을 때 보유IPC수와 보유CPC수의 상관관계가 0.723으로 다중공선성을 의심해볼 만한 수치가 도출되었다.

	보유IPC수	보유CPC수	도면수	발명자수
보유IPC수	1.000000	0.722869	0.024172	0.165586
보유CPC수	0.722869	1.000000	0.220712	0.114167
도면수	0.024172	0.220712	1.000000	0.100085
발명자수	0.165586	0.114167	0.100085	1.000000

독립변수간의 상관계수

통상적으로 사용되는 독립변수는 보유IPC수이고 보유CPC수는 우리가 임의로 추가한 변수였다. 따라서 보유CPC수와 특허승인소요기관 변수를 제외하고 다시 다중회귀분석을 시도해보았다.

12) 다중회귀분석 3번째 시도

OLS Regression Results

Dep. Variable:	특허의후방인용수	**R-squared:**	0.109
Model:	OLS	**Adj. R-squared:**	0.109
Method:	Least Squares	**F-statistic:**	1540.
Date:	Thu, 31 Mar 2022	**Prob (F-statistic):**	0.00
Time:	03:17:56	**Log-Likelihood:**	-2.6620e+05
No. Observations:	37580	**AIC:**	5.324e+05
Df Residuals:	37576	**BIC:**	5.324e+05
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	12.8697	3.061	4.204	0.000	6.869	18.870
보유IPC수	0.6789	0.313	2.166	0.030	0.064	1.293
도면수	3.4005	0.051	66.604	0.000	3.300	3.501
발명자수	3.8885	0.636	6.113	0.000	2.642	5.135

Omnibus:	55495.093	**Durbin-Watson:**	1.578
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	23223015.362
Skew:	9.094	**Prob(JB):**	0.00
Kurtosis:	123.417	**Cond. No.**	73.4

보유CPC수 변수를 제외한 결과, 오히려 더 낮은 수정 결정계수가 도출됨을 확인할 수 있었으며 다중 회귀분석기법을 이용한 위생학 분야 특허의 피인용 수 예측 회귀식은 다음과 같이 나타낼 수 있다.

$$Y = 5.2616 - 5.5941*(\text{보유}IPC\text{수}) + 5.5640*(\text{보유}CPC\text{수}) + 3.0898*(\text{도면수}) + 4.3712*(\text{발명자수})$$

V. 결론 및 한계점

본 보고에서는 위생학 분야의 특허 피 인용수를 예측하기 위한 분석을 수행하였다. 주어진 데이터에서 사용할 수 있는 독립변수는 기존에 선행적으로 연구되던 피 인용수를 예측하기 위해 통상적으로 사용되는 독립변수에서 많이 추출할 수 없었다. 제시된 데이터에서 '보유 CPC 수', '보유 IPC 수', '도면수', '발명자수', '특허승인소요기간'을 독립변수로 선정하여 다중회귀분석을 실시한 결과로, 결정계수와 수정된 결정계수의 값이 0.119 로 매우 설명력이 낮은 결과가 도출되었다. 추가적으로 '특허승인소요기간' 변수의 p-value 값이 0.099 로, 통계적으로 유의미하지 못한다는 결론을 내릴 수 있었다. 따라서 변수 선택법으로 전진 선택법, 후진 제거법, 단계적 선택법을 사용하였고 모든 변수 선택법의 결과로, '특허승인소요기간'을 제외한 나머지 4 가지 독립변수를 사용하는 것이 좋은 값을 도출하는 것으로 확인되었다.

따라서, '특허승인소요기간' 독립변수를 제외한 나머지 4 가지의 독립변수를 사용하여 다중회귀분석을 다시 시행한 결과, 마찬가지로 수정 결정계수가 0.119 라는 값이 도출되었다. 다중공선성에 문제가 있을 수도 있다고 판단하였기에, 분산팽창인자를 도출하여 다중공선성을 확인해보았지만 이에도 문제가 없었고, '보유 IPC 수'와 '보유 CPC 수'의 상관관계가 0.7 이상으로 강한 양의 상관관계를 띄어, '보유 CPC 수' 변수를 제외하고 다시 다중회귀분석을 시행하였지만, 수정 결정계수는 0.109 로 더 낮은 설명력을 보여주는 결과가 도출되었다.

다중회귀분석의 결과로 수정 결정계수가 0.119 로 낮은 설명력을 보여주지만, F-통계량이 0.00 으로 통계적으로 유의하다는 결과가 도출되었고 Durbin-Watson 의 값이 1 보다 크므로, 종속변수와 독립변수 사이에 자기 상관관계가 없음을 확인할 수 있었다.

위생학 분야의 특허 피 인용수를 예측하기 위한 회귀식은 다음과 같다.

$$Y = 5.2616 - 5.5941*(\text{보유}IPC\text{수}) + 5.5640*(\text{보유}CPC\text{수}) + 3.0898*(\text{도면수}) + 4.3712*(\text{발명자수})$$

분야마다의 보유 IPC 수, 보유 CPC 수, 발명자수에 따라 특허의 피 인용수에 차이가 존재할 것이므로, 각 분야마다의 회귀식과 수정 설정계수는 전부 다르게 도출될 것이다. 더불어, 수정 결정계수가 낮으므로 회귀분석에 사용된 독립변수들이 특허의 피 인용수에 확연한 어떤 관계를 가진다고 할 수 없을 것이다. 수정 결정계수가 낮게 도출된 이유에 대하여, 너무 적은 독립변수의 수를 들 수 있을 것이다. 따라서 다른 독립변수들이 추가적으로 주어진다면, 위에서 제시하였던 특허의 피 인용수에 영향을 끼치는 더 다양한 변수를 도출해낼 수 있을 것이며, 그에 따라 더 높은 결정계수를 띄는 회귀식을 도출해낼 수 있다고 예상할 수 있다.

References

[1] 조현진, & 이학연. (2018). 머신러닝 기법을 활용한 특허 품질 예측. *대한산업공학회 춘계공동학술대회 논문집*, 1343-1350.

[2] Yoo-jae Lee, "A Study on the Verification of the Main Effect in Multiple Regression Analysis including Interaction Effect", Management Research, Vol. 23, No. 4, pp. 183-210, 1994.7.

[3] Uh-Soo Kyun, Sung-Hoon Cho, Jeong-Joon Kim, “A Study on Perception for Public Safety of Seoul Citizens using Multiple Regression Analysis”, The Journal of The Institute of Internet, Broadcasting and Communication, Vol. 18, No. 1, pp. 195-201, Feb 2018. DOI: https://doi.org/10.7236/JIIBC.2018.18.1.195

[4] Bong-Woo Nam, Kyung-Bin Kim, Kyu-Ho Kim, Jun-Min Cha, "Regional Power Demand Forecasting Algorithm Using Multiple Regression Analysis", Journal of the Korean Institute of Illuminating and Electric

[5] 임준묵.(2020).대학생의 중도탈락의도에 영향을 미치는 교육 요인.한국엔터테인먼트산업학회논문지,14(3),105-115.