당뇨병 진행에 관한 모델링 분석

2021190002 장서현

I. 서론

본 분석 보고서는 당뇨병 진행 상태에 영향을 미치는 주요 요인들을 식별하고, 그 요인들을 중심으로 모델링을 진행함으로써 당뇨병 진행 상태를 얼마나 정확하게 예측할 수 있는지를 분석하는 것을 의의로 한다. 분석에 사용된 데이터셋("diabetes.cvs")은 나이, 혈압, 인슐린 수치, BMI와 같은 다양한 생리학적 측정값들과 당뇨병 진행 상태를 측정한 결과를 포함하고 있다. 이러한 변수들이 당뇨병 진행 상태를 이해하고 예측하는 데 있어 미치는 각각의 영향력을 면밀히 조사하고, 이러한 변수들 간의 관계를 파악하고자 한다.

II. 분석 내용

주어진 모든 변수를 이용한 모델을 '모델1'이라 칭하고 모델1의 결과를 분석하자. 데이터를 활용하여 학습시키고 OLS Regression으로모델링한 결과, 모델1은 Adjusted R-squared은 0.5587의 값이 나오고, Mean Squared Error (이하 MSE)는 2772.914의 값이 나온다.

1) P-Value 분석: 변수 'age', 's3', 's4', 's6'이 p-value의 값이 높게 나타나 해당 변수들이 통계적으로 유의미하지 않을 가능성을 염두 하여야 한다. (Appendix A 참고)

2) 다중공선성 (multicollinearity) 분석:

   A. Heatmap: 히트맵에서 색상은 변수 간의 관계의 강도를 나타낸다. 진한색은 강한 상관관계를 나타내고, 밝은 색은 약한 상관관계를 나타낸다. 히트맵의 대각선은 각 변수와 자기 자신의 상관관계를 나타내므로 완벽한 상관관계를 이룬다. 이를 제외하고, 변수 's1'와 's2' 그리고 변수 's4'와 's5'는 색이 진한 것으로 보아 다중 공선성을 의심할 수 있다. (Appendix B 참고)

   B. Pair-plot: 페어플롯은 변수 쌍의 산점도를 보여준다. 변수 's1'와 's2' 사이에 선형 관계와 그에 따른 다중공선성이 있을 것으로 예상할 수 있다. 다만, 히트맵과 페어플롯은 수치로서 두 변수의 다중공선성을 명확히 명시하지 못하므로 VIF를 계산함으로써 확인한다. (Appendix C 참고)

   C. VIF Factor: VIF는 다중 회귀 모형에서 변수 간의 상관관계를 측정하는 인덱스이다. 변수 's1'은 59.063의 값, 변수 's2'는 39.123의 값으로 두 변수 사이에 다중 공선성이 있음을 알 수 있다. (Appendix D 참고)

히트맵, 페어플롯, VIF 값을 바탕으로 두 변수 's1'과 's2' 사이에 상관관계가 있음을 알 수 있고, 다중공선성 강하다는 것 또한 알 수 있다. 's1'은 T-cell 수치, 's2'는 low density lipoproteins

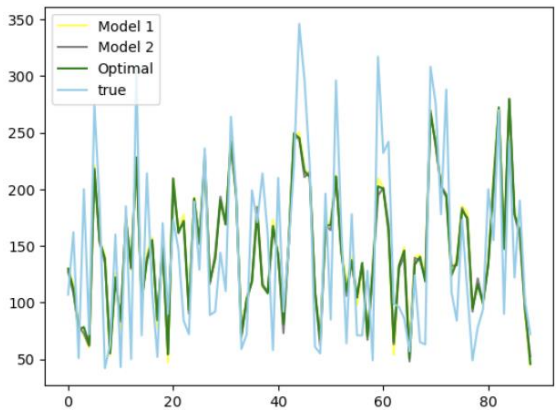
수치를 나타낸다. 당뇨병은 인슐린 저항성과 관련이 깊은데, 인슐린 저항성은 LDL 콜레스테롤의 증가와 연관이 있고, 동시에 염증 반응을 촉진 시켜 T세포의 활성화를 증가시킬 수 있다. 이 뿐만 아닌, 다양한 생리학적 근거로 두 변수 사이에 상관관계와 다중공성성이 발생할 수 있다.

과적합의 위험을 증가시키는 다중 공선성을 없애기 위해 변수 's2' 제거한 뒤, 이를 '모델2'라 칭하자. 데이터를 활용하여 학습시키고 OLS Regression으로 모델링한 결과, 모델2은 Adjusted R-squared은 0.5626의 값으로 '모델1'에 비해 피팅이 더 잘 된 것을 관찰할 수 있고, MSE는 2748.761의 값으로 '모델1'에 비해 에러가 적게 나온다. 's2'를 제거한 결과 VIF 값 도출 결과 '모델2'에 다중 공선성이 없어짐을 알 수 있다 (Appendix E 참고). '모델2'의 OLS Regression Summary를 보면, 변수 'age', 's3', 's4'의 p-value가 0.05 수준에서 통계적으로 유의미하지 않다는 것이 확인된다. 변수 's6' 또한 상대적으로 높은 p-value를 가지고 있는 것을 보아, 모델링에 방해가 되는 변수일 수 있다 (Appendix F 참고).

다중 공선성을 띄지 않으며, MSE가 낮고, Adjusted R-squared가 높은 모델을 찾는 것을 목표로 '모델2'에서 p-value가 높았던 변수들을 제거하는 방향으로 다양한 조합의 모델링을 시도하였다. 그 결과, 변수 'age', 's2', 's3', 's6'를 제거한 모델이 가장 피팅이 잘 되었으며, 이 모델을 'Optimal Model'이라 칭하자. 'Optimal Model'의 MSE 값이 2702.7322, 그리고 Adjusted R-Squared 값이 0.570으로 나머지 모델들에 비해 MSE 값이 낮고, Adjusted R-squared 값이 높다. 'Optimal Model'의 OLS Regression Summary를 확인하면, 모든 변수가 유의미하다는 것 또한 확인할 수 있다 (Appendix G 참고). 마지막으로, VIF factor로 변수들 사이의 상관관계를 측정해보면, 'Optimal Model'에는 다중 공선성이 없다는 것도 확인 가능하다 (Appendix H 참고).

III. 결론

본 분석을 통해, 현재 구축한 'Optimal Model'이 당뇨병 진행 상태를 충분히 설명하지 못하고 있는 것으로 나타난다. Adjusted R-squared가 낮은 결과는 모델이 데이터 내의 변동을 효과적으로 포착하지 못하고 있음을 시사하는데, 이는 다중 공선성을 해결하고 통계적으로 유의미하지 않은 변수들을 제거한 'Optimal Model'에서도 마찬가지이다.



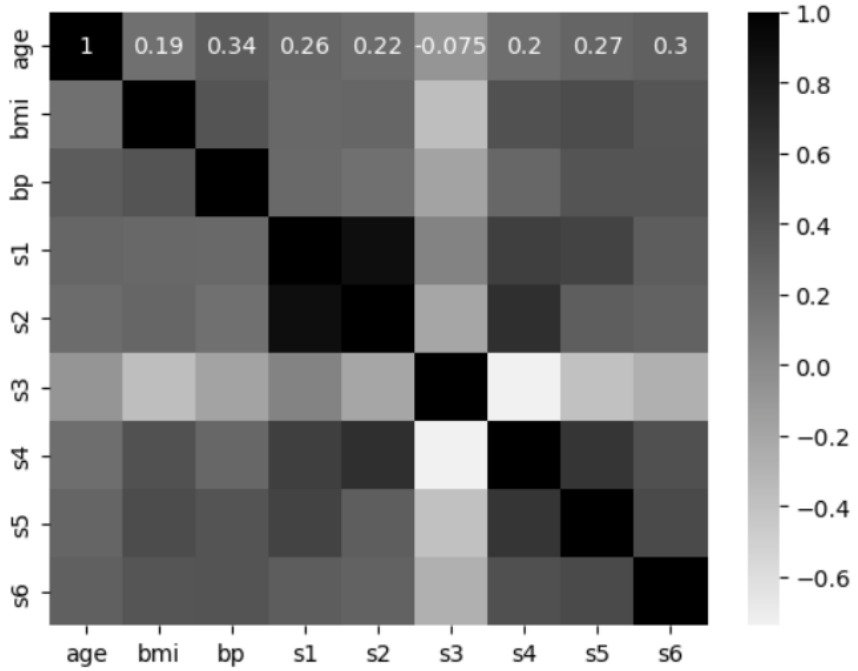
위 그래프는 '모델1', '모델2', 'Optimal Model'을 OLS Regression을 이용해 학습시켜 모델링한 결과물을 시각적으로 보여준다. 'Optimal Model'의 Adjusted R-squared가 0.570로 낮은 수치이며, 그에 따라 'Optimal Model'은 'True Model'을 엄밀히 설명하고 있지 못하다. 이에 따라, 모델의 예측력을 개선하기 위해서는 추가적인 변수의 도입이 필요하다고 판단된다. 예를 들면, 가족력의 여부, 신체 활동 수준, 흡연 여부와 같은 변수들은 당뇨병의 진행과 밀접한 관련이 있으며, 모델에 포함될 경우 설명력을 높이는 데 기여할 수 있을 것으로 예상된다.

또한, 적절한 샘플 사이즈의 확보는 모델의 정확도와 일반화 가능성을 높이는 데 중요하다. 샘플 사이즈가 작은 경우, 모델이 데이터의 특정 패턴을 과대적합 할 위험이 있으며, 이는 새로운 데이터에 대한 예측력을 저하시킬 수 있다. 본 데이터셋은 442개의 샘플만 있는 것으로 보아, 더 많은 데이터를 수집하고 분석에 포함시키는 것은 모델의 피팅을 개선하고, 당뇨병 진행 상태 예측의 정확도를 높이는 데 도움이 될 것으로 예상된다.

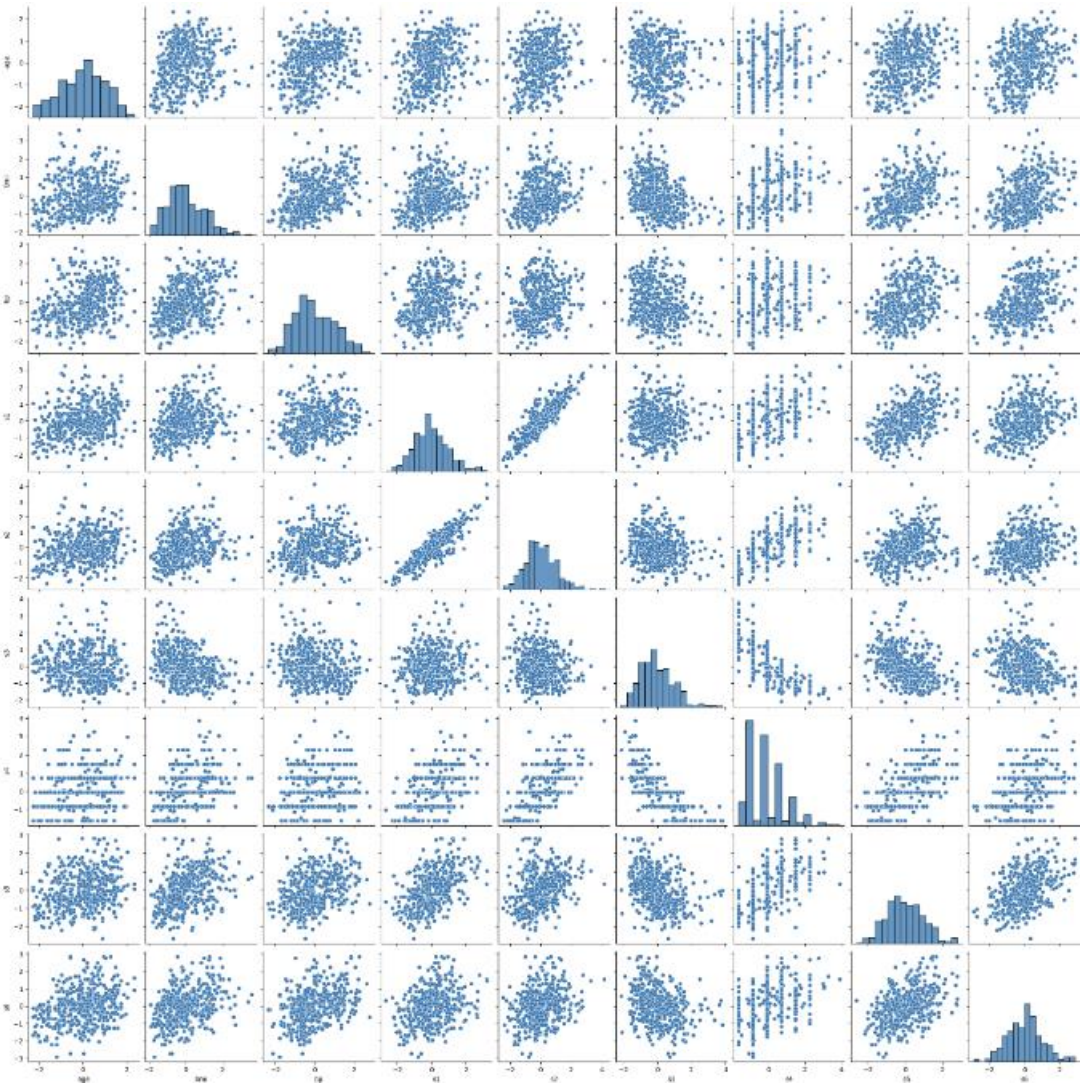
Appendix A. Model 1 – OLS Regression Summary

	coef	std err	t	P>|t|	[0.025	0.975]
const	152.1335	2.618	58.105	0.000	146.987	157.280
age	-1.5780	2.875	-0.549	0.583	-7.228	4.072
bmi	26.4967	3.184	8.322	0.000	20.239	32.754
bp	13.1319	3.106	4.227	0.000	7.026	19.237
s1	-33.9046	20.122	-1.685	0.093	-73.454	5.644
s2	20.0042	16.377	1.221	0.223	-12.184	52.193
s3	6.6359	10.265	0.646	0.518	-13.539	26.811
s4	6.0065	7.782	0.772	0.441	-9.289	21.302
s5	35.9767	8.311	4.329	0.000	19.642	52.311
s6	2.3268	3.182	0.731	0.465	-3.927	8.581

Appendix B. Heatmap of Model 1



Appendix C. Pair-plot of Model 1



Appendix D. VIF factors of Model 1

VIF Factor	features
59.062508	s1
39.123245	s2
15.369272	s3
10.075391	s5
8.833675	s4
1.478660	bmi
1.476845	s6
1.407578	bp
1.205380	age

Appendix E. VIF factors after removing variable 's2'

VIF Factor	features
8.676686	s4
6.468346	s3
4.416862	s1
2.071229	s5
1.476304	s6
1.467012	bmi
1.407234	bp
1.202814	age

Appendix F. Model 2 – OLS Regression Summary

	coef	std err	t	P>|t|	[0.025	0.975]
const	152.1335	2.620	58.072	0.000	146.984	157.283
age	-1.4160	2.873	-0.493	0.622	-7.063	4.231
bmi	26.8418	3.173	8.459	0.000	20.605	33.078
bp	13.0725	3.108	4.206	0.000	6.964	19.181
s1	-10.2627	5.506	-1.864	0.063	-21.084	0.559
s3	-2.9058	6.663	-0.436	0.663	-16.001	10.190
s4	4.7394	7.717	0.614	0.539	-10.428	19.906
s5	26.9285	3.770	7.142	0.000	19.518	34.339
s6	2.4012	3.183	0.754	0.451	-3.855	8.657

Appendix G. Optimal Model – OLS Regression Summary

	coef	std err	t	P>|t|	[0.025	0.975]
const	152.1335	2.613	58.212	0.000	146.997	157.270
bmi	27.3828	3.097	8.842	0.000	21.296	33.469
bp	13.2503	2.957	4.481	0.000	7.438	19.062
s1	-12.2710	3.242	-3.785	0.000	-18.643	-5.899
s4	8.0760	3.601	2.243	0.025	0.999	15.153
s5	27.4813	3.675	7.478	0.000	20.258	34.704

Appendix H. VIF Factors of Optimal Model

VIF Factor	features
1.977411	s5
1.898093	s4
1.538962	s1
1.404091	bmi
1.280268	bp