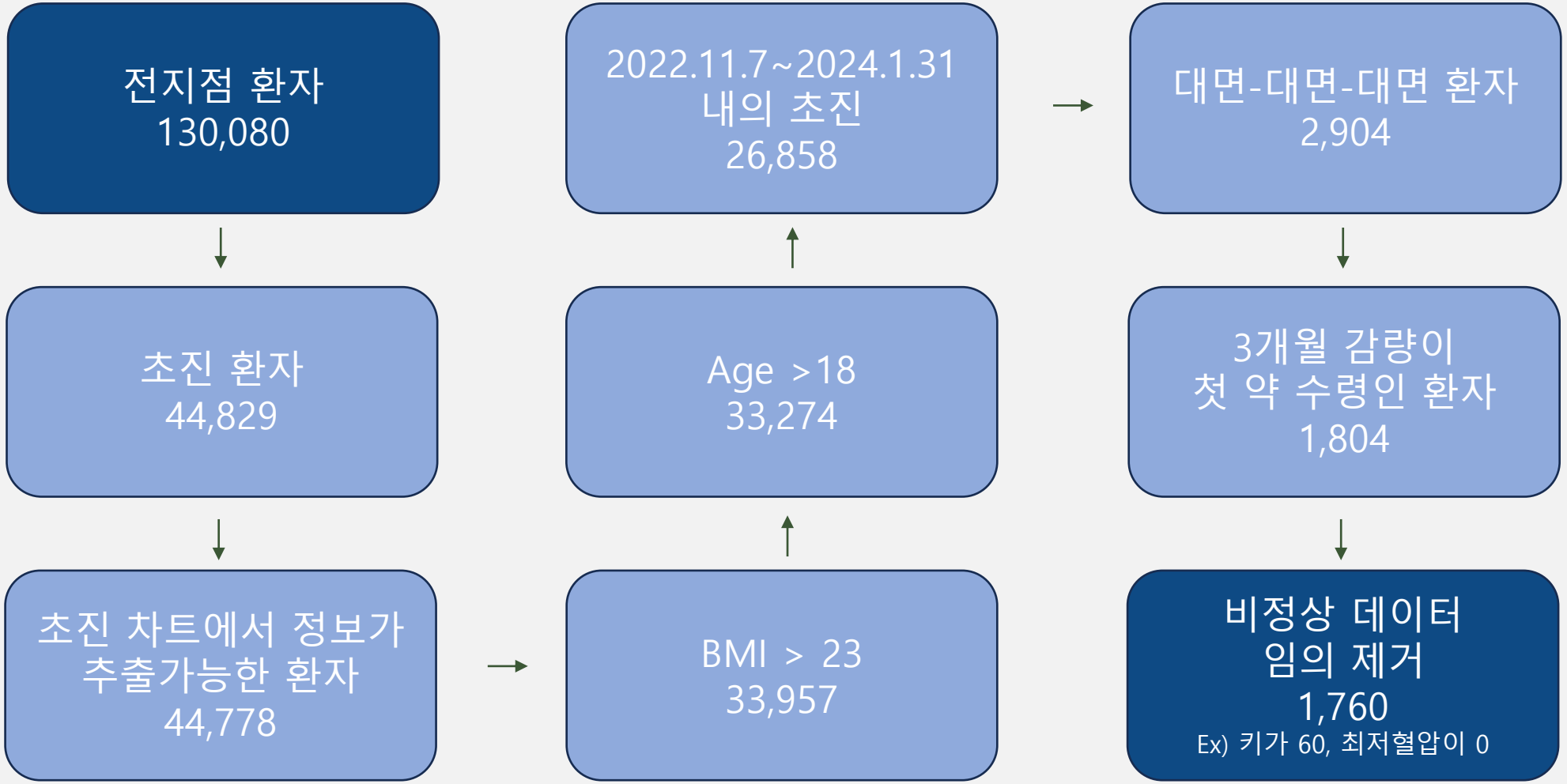
다중선형 회귀 모델을 이용한 3개월 감량 예측 모델

누베베 인턴 정성준

목차

I 데이터

데이터 추출과정



I 데이터

수치형 독립변수 소개

변수명	평균	표준편차	최소값	중간값	최대값
Age	39.87	11.66	19	39	71
Height	162.86	7.20	130	162	190
Weight	75.75	13.65	49.90	72.50	149.30
BMI	28.58	3.95	21.00	27.85	48.90
BMR	1366	190	1032	1320	2453
FatFreeMass	46.14	8.78	30.70	44.00	96.4
PBF	38.88	5.73	18.30	38.90	53.30
SMM	25.00	5.32	16.00	23.90	55.80

I 데이터

수치형 독립변수 소개

변수명	평균	표준편차	최소값	중간값	최대값
SoftLeanMass	43.44	8.30	29.00	41.35	90.90
VFA	144.85	39.13	51.70	140.80	280.70
WHR	0.93	0.06	0.76	0.92	1.22
apedrin1	109.04	11.18	80	105	145
apedrin2	115.8	11.72	80	115	145
Period	90.58	14.29	60	91	120
MaxVital_1	134.50	17.20	88	134	206
MinVital_1	81.42	11.75	38	81	131
Pulse_1	84.83	12.69	33	83	133

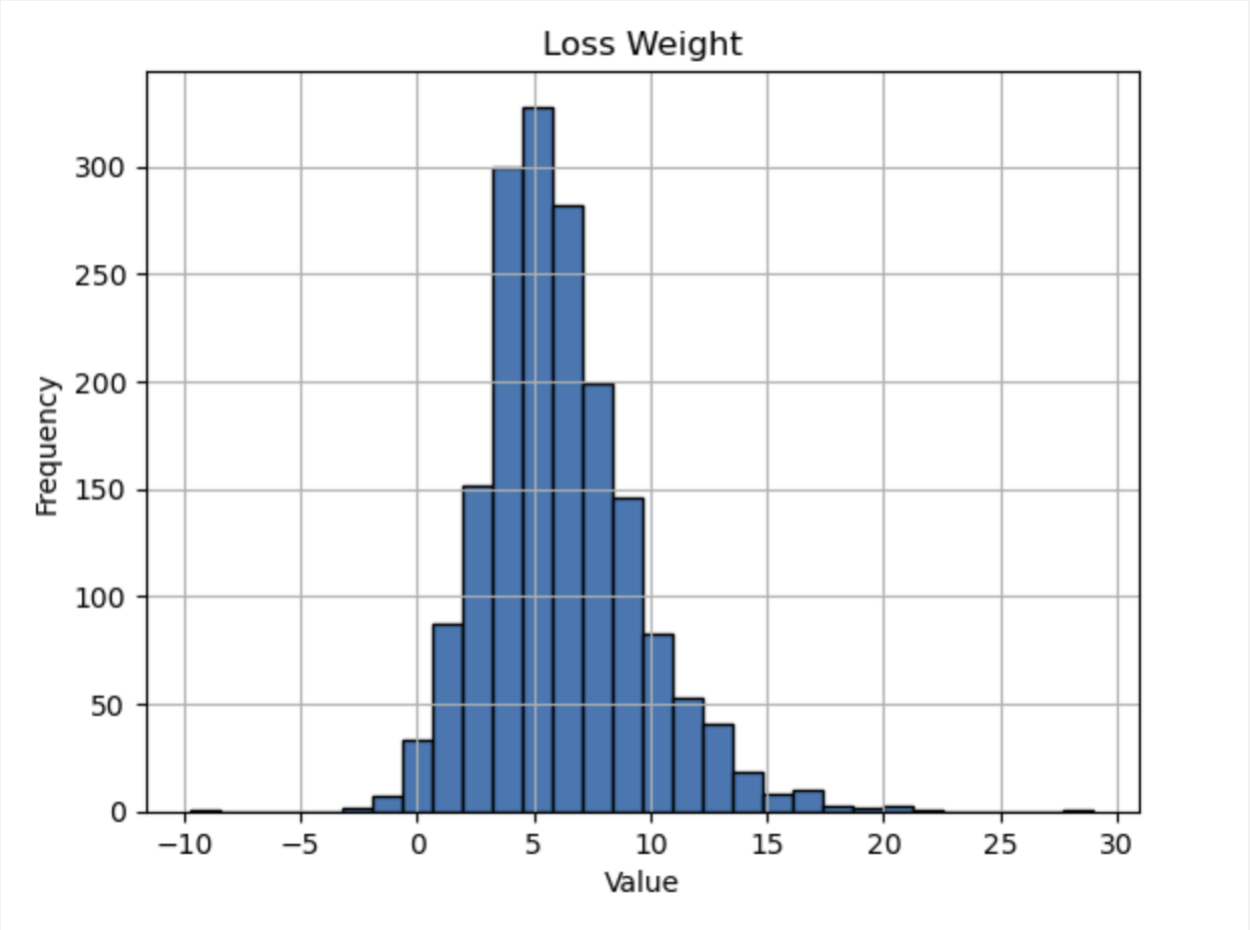
I 데이터

범주형 독립변수 소개

변수명		변수명	
PatientSex	1:남자, 2:여자	Coffee_기타	0, 1
HanbangX 한방 경험 없음	0, 1	Coffee_마시지 않음	0, 1
HanbangYX 한방 경험 있지만 불편증상 없음	0, 1	Coffee_없음	0, 1
HanbangYY 한방 경험 있고 불편증상 있음	0, 1	Coffee_있음	0, 1
YangbangX 양방 경험 없음	0, 1	Alcohol_거의 마시지 않는다	0, 1
YangbangYX 양방경험 있지만 불편증상 없음	0, 1	Alcohol_주 1회	0, 1
YangbangYY 양방 경험 있고 불편증상 있음	0, 1	Alcohol_주 2회 이상	0, 1
		Alcohol_기타	0, 1

Ⅱ EDA

3개월감량



- 정규분포 형태
- 0~15 사이에 대부분의 사람이 몰려 있음

II EDA

3개월 감량과의 상관계수 (양의 상관관계)

Weight_1	0.490783
BMI_1	0.459807
VFA_1	0.418728
WHR_1	0.343075
FatFreeMass_1	0.340533
BMR_1	0.340358
SoftLeanMass_1	0.339439
SMM_1	0.336669
apedrin1	0.285372
MaxVital_1	0.244290
PBF_1	0.237170
MinVital_1	0.216909
Height	0.213465
apedrin2	0.167051
YangbangX	0.138544
Pulse_1	0.138397
HanbangX	0.098418
Period	0.094016
Alcohol_거의 마시지 않는다 (월 1-2회)	0.040984
Coffee_없음	0.025776
Coffee_마시지않음	0.018314

- Weight, BMI, VFA 순으로 상관관계가 큼

II EDA

체중과의 상관계수 (양의 상관관계)

BMR_1	0.848236
FatFreeMass_1	0.848198
SoftLeanMass_1	0.847854
BMI_1	0.846391
SMM_1	0.843210
apedrin1	0.685551
WHR_1	0.673641
VFA_1	0.652267
apedrin2	0.588120
Height	0.585541
MaxVital_1	0.457389
MinVital_1	0.386878
Pulse_1	0.204070
PBF_1	0.203380

- 감량과 높은 상관계수를 보이는 대부분의 변수들이 체중과 상관관계가 크다
- 다중공선성을 고려하여 높은 변수만 활용하는건 좋은 방법이 아닌 것으로 보임

3개월 감량과의 상관계수 (음의 상관관계)

Alcohol_주 2회 이상	-0.008555
Coffee_기타	-0.017489
Coffee_있음	-0.033244
HanbangYY	-0.039731
Alcohol_주 1회	-0.040611
HanbangYX	-0.056877
YangbangYX	-0.070580
YangbangYY	-0.094700
PatientSex	-0.179358
Age	-0.231971

- 뚜렷한 음의 상관 관계를 보이는 변수는 확인 되지 않는다.

Ⅲ 변수선택

변수선택 전략

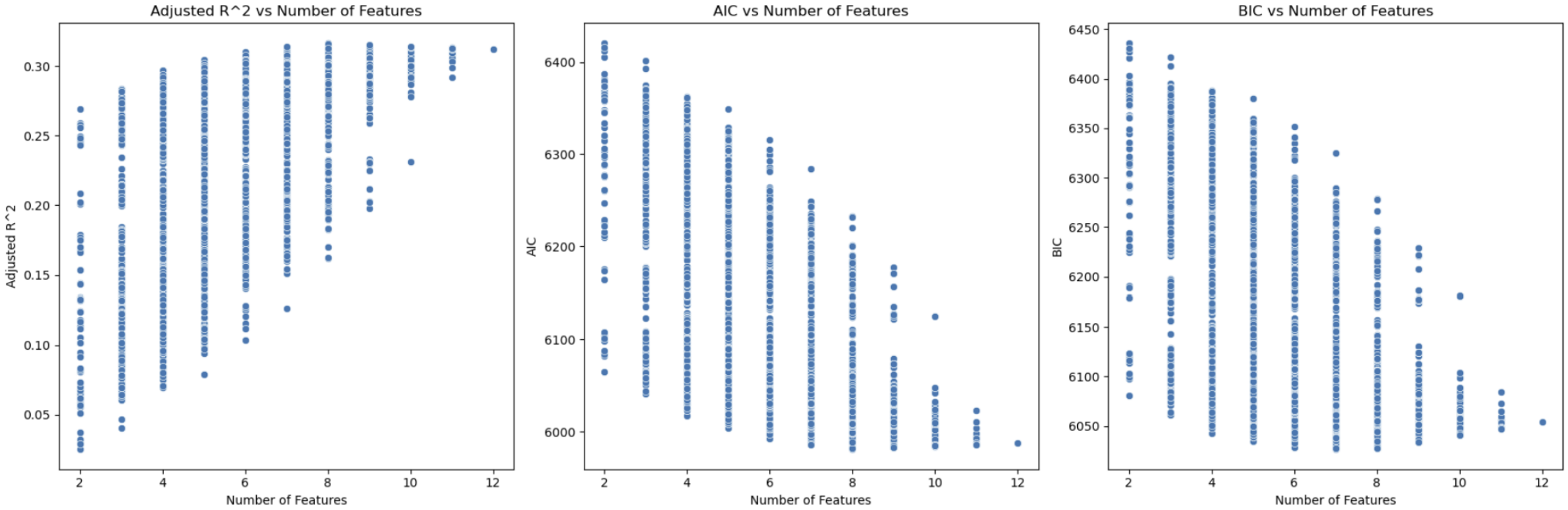
- 32개의 변수로 가능한 조합은 2^{32}으로 모든 변수를 테스트 하는 것은 무리가 있음
- Lasso를 활용하여 alpha 값을 조절하여 15개 이하의 변수를 1차적으로 추출
- 1차 검열된 변수들로 가능한 모든 조합을 평가하여 $Adj\ R^2, AIC, BIC$로 각각 평가하여 변수조합을 찾음

Lasso

- Alpha = 0.1 에서 12개의 변수를 남기고 탈락
- Age, Height, YangbangX, BMR, MaxVital, MinVital, Pulse, VFA, Weight, apedrin1, apedrin2, Period

Ⅲ 변수선택

Grid search




```
Best R^2 Combination: ('Age', 'YangbangX', 'BMR_1', 'MinVital_1', 'Weight_1', 'apedrin1', 'apedrin2', 'Period')
Best AIC Combination: ('Age', 'YangbangX', 'BMR_1', 'MinVital_1', 'Weight_1', 'apedrin1', 'apedrin2', 'Period')
Best BIC Combination: ('Age', 'YangbangX', 'BMR_1', 'Weight_1', 'apedrin1', 'apedrin2', 'Period')
```

- R^2와 AIC에서 나온 조합이 겹치기 때문에 1차적으로 1,2번 조합을 선택

Ⅳ 다중 회귀 적합

적합 결과

OLS Regression Results

Dep. Variable:	Weight_After	R-squared:	0.297
Model:	OLS	Adj. R-squared:	0.292
Method:	Least Squares	F-statistic:	64.46
Date:	Tue, 06 Aug 2024	Prob (F-statistic):	3.78e-88
Time:	12:20:19	Log-Likelihood:	-3009.9
No. Observations:	1232	AIC:	6038.
Df Residuals:	1223	BIC:	6084.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.2948	1.128	-1.148	0.251	-3.508	0.919
Age	-0.0343	0.007	-4.638	0.000	-0.049	-0.020
YangbangX	0.7544	0.178	4.248	0.000	0.406	1.103
MaxVital_1	0.0137	0.006	2.491	0.013	0.003	0.025
VFA_1	0.0133	0.003	4.825	0.000	0.008	0.019
Weight_1	0.0831	0.010	8.040	0.000	0.063	0.103
apedrin1	0.0730	0.018	4.107	0.000	0.038	0.108
apedrin2	-0.0971	0.015	-6.540	0.000	-0.126	-0.068
Period	0.0164	0.006	2.909	0.004	0.005	0.027

Omnibus:	111.275	Durbin-Watson:	1.938
Prob(Omnibus):	0.000	Jarque-Bera (JB):	387.790
Skew:	0.397	Prob(JB):	6.20e-85
Kurtosis:	5.632	Cond. No.	4.03e+03

- 모든 계수가 유의수준 0.05이하의 p-value 값을 가짐
- 수정설명계수 값이 0.322로 높다고 보기 힘듬

IV 다중 회귀 적합

개선방안

OLS Regression Results

Dep. Variable:	Weight_After	R-squared:	0.305
Model:	OLS	Adj. R-squared:	0.300
Method:	Least Squares	F-statistic:	59.66
Date:	Tue, 06 Aug 2024	Prob (F-statistic):	1.74e-90
Time:	12:20:19	Log-Likelihood:	-3002.2
No. Observations:	1232	AIC:	6024.
Df Residuals:	1222	BIC:	6076.
Df Model:	9		
Covariance Type:	nonrobust		

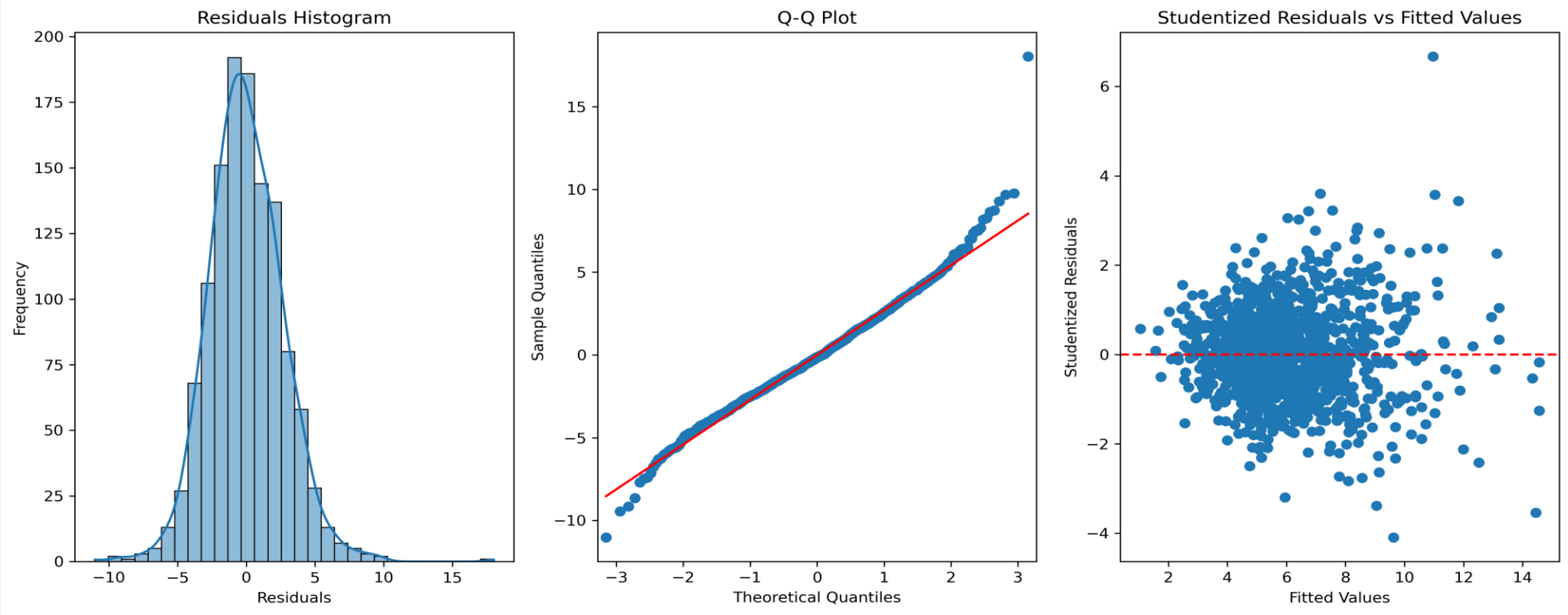
	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.6534	1.134	-0.576	0.564	-2.878	1.571
Age	-0.0356	0.007	-4.841	0.000	-0.050	-0.021
YangbangX	0.7479	0.177	4.235	0.000	0.401	1.094
MaxVital_1	0.0145	0.005	2.640	0.008	0.004	0.025
VFA_1	0.0130	0.003	4.751	0.000	0.008	0.018
Weight_1	0.0837	0.010	8.140	0.000	0.063	0.104
apedrin1	0.0686	0.018	3.874	0.000	0.034	0.103
apedrin2	-0.0950	0.015	-6.434	0.000	-0.124	-0.066
Period	0.0163	0.006	2.911	0.004	0.005	0.027
absPeriod	-0.0377	0.010	-3.910	0.000	-0.057	-0.019

Omnibus:	115.968	Durbin-Watson:	1.945
Prob(Omnibus):	0.000	Jarque-Bera (JB):	433.683
Skew:	0.394	Prob(JB):	6.71e-95
Kurtosis:	5.798	Cond. No.	4.07e+03

- 수정 설명 계수를 올리기 위해 변수들의 편차의 제곱항($(x - mean(x))^2$)을 시도
- Period의 편차 제곱항 을 추가 했을 경우만 계수가 유의 한 결과가 나옴
- 수정계수 약간 상승
- 비슷한 의미를 가지는 편자의 절대값 항 ($|x - mean(x)|$) 을 시도
- 수정계수가 약간 더 상승
- 최종적으로 Period의 편차 절대값항을 추가하기로 결정

V 회귀 진단

잔차분석



- 잔차들이 정규분포에서 벗어나며 이상치로 판단되는 점들이 보임
- 일반적인 기준을 적용하여 준스튜던트화 잔차가 3 이상의 값을 이상치로 판단하고 제거
- 10개의 점 재거

V 회귀 진단

모델 재적합

OLS Regression Results

Dep. Variable:	Weight_After	R-squared:	0.319
Model:	OLS	Adj. R-squared:	0.314
Method:	Least Squares	F-statistic:	63.10
Date:	Tue, 06 Aug 2024	Prob (F-statistic):	6.47e-95
Time:	12:22:19	Log-Likelihood:	-2889.0
No. Observations:	1222	AIC:	5798.
Df Residuals:	1212	BIC:	5849.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1381	1.061	0.130	0.896	-1.944	2.220
Age	-0.0351	0.007	-5.117	0.000	-0.049	-0.022
YangbangX	0.7184	0.165	4.359	0.000	0.395	1.042
MaxVital_1	0.0098	0.005	1.913	0.056	-0.000	0.020
VFA_1	0.0135	0.003	5.254	0.000	0.008	0.019
Weight_1	0.0859	0.010	8.818	0.000	0.067	0.105
apedrin1	0.0522	0.017	3.147	0.002	0.020	0.085
apedrin2	-0.0811	0.014	-5.881	0.000	-0.108	-0.054
Period	0.0138	0.005	2.642	0.008	0.004	0.024
absPeriod	-0.0405	0.009	-4.501	0.000	-0.058	-0.023

Omnibus:	17.115	Durbin-Watson:	1.989
Prob(Omnibus):	0.000	Jarque-Bera (JB):	17.539
Skew:	0.274	Prob(JB):	0.000155
Kurtosis:	3.211	Cond. No.	4.08e+03

- 성능이 향상된 것을 확인
- 하지만 MaxVital의 p-value가 유의수준 0.05 보다 큼

V 회귀 진단

모델 재적합

OLS Regression Results

Dep. Variable:	Weight_After	R-squared:	0.317
Model:	OLS	Adj. R-squared:	0.312
Method:	Least Squares	F-statistic:	70.37
Date:	Tue, 06 Aug 2024	Prob (F-statistic):	4.54e-95
Time:	13:48:20	Log-Likelihood:	-2890.8
No. Observations:	1222	AIC:	5800.
Df Residuals:	1213	BIC:	5846.
Df Model:	8		
Covariance Type:	nonrobust		

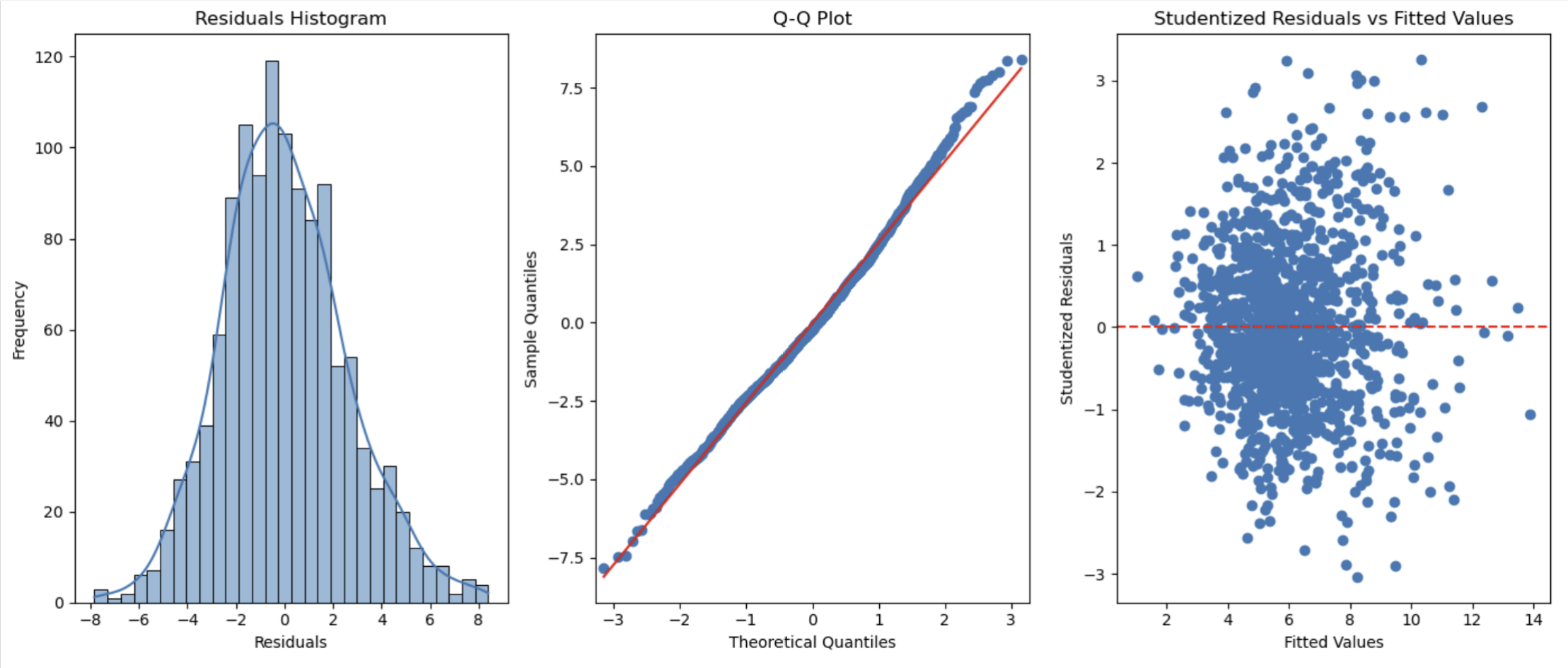
	coef	std err	t	P>|t|	[0.025	0.975]
const	0.8109	1.002	0.809	0.419	-1.156	2.777
Age	-0.0321	0.007	-4.802	0.000	-0.045	-0.019
YangbangX	0.7430	0.164	4.517	0.000	0.420	1.066
VFA_1	0.0140	0.003	5.475	0.000	0.009	0.019
Weight_1	0.0911	0.009	9.707	0.000	0.073	0.109
apedrin1	0.0526	0.017	3.172	0.002	0.020	0.085
apedrin2	-0.0808	0.014	-5.855	0.000	-0.108	-0.054
Period	0.0133	0.005	2.553	0.011	0.003	0.024
absPeriod	-0.0399	0.009	-4.434	0.000	-0.058	-0.022

Omnibus:	19.574	Durbin-Watson:	1.991
Prob(Omnibus):	0.000	Jarque-Bera (JB):	20.294
Skew:	0.290	Prob(JB):	3.92e-05
Kurtosis:	3.249	Cond. No.	3.39e+03

- Maxvital 제거
- 모든 계수가 유의 수준보다 작음

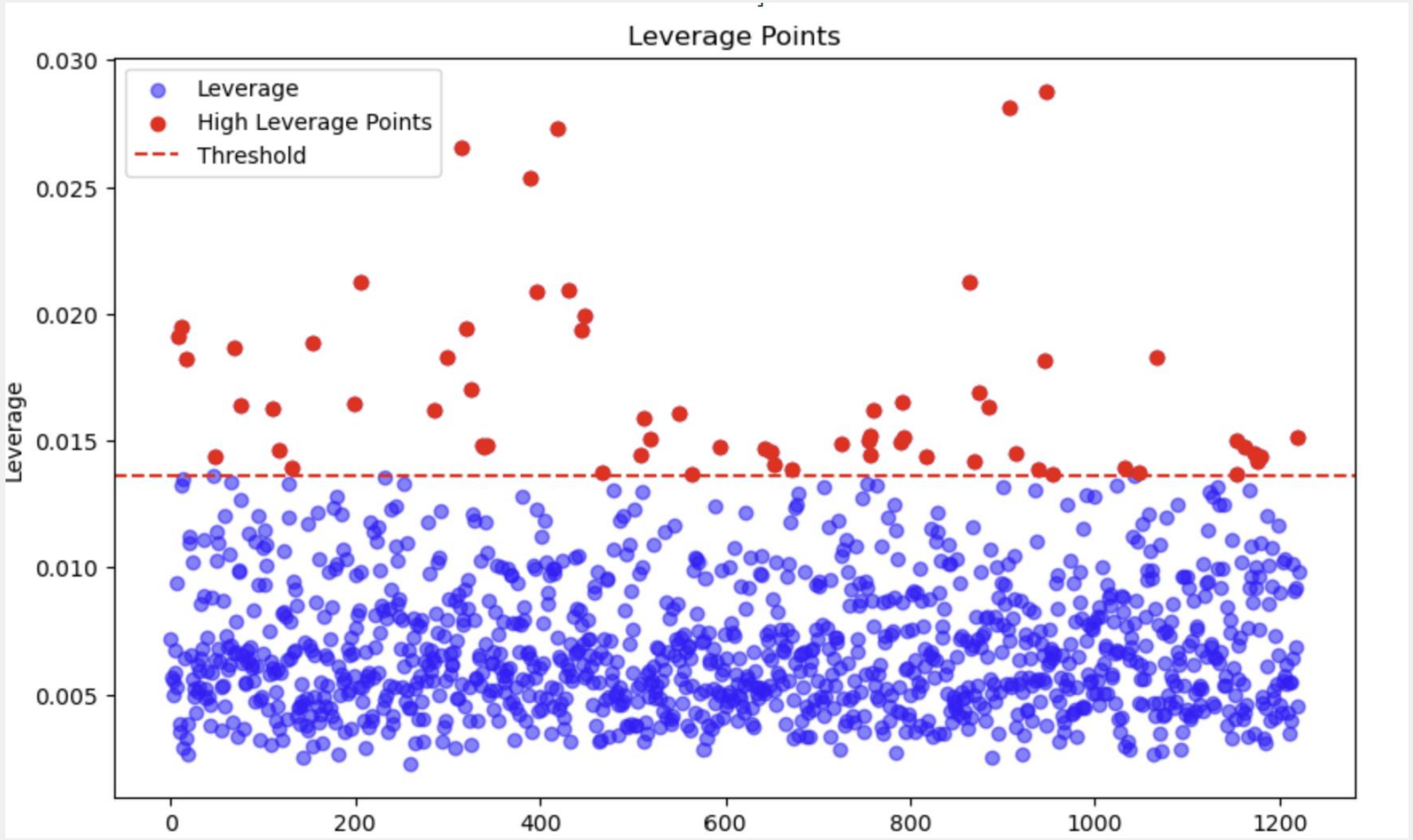
V 회귀 진단

잔차 재분석


Residuals Histogram
Frequency
Residuals
Q-Q Plot
Sample Quantiles
Theoretical Quantiles
Studentized Residuals vs Fitted Values
Studentized Residuals
Fitted Values

V 회귀 진단

High Leverage Points



- 영향력이 평균보다 2배 이상 높은 데이터가 많음 (66개)
- High Leverage Points를 제거하고 회귀 적합 하여도 성능이 좋아지지 않음
- 빼지 않기로 결정

V 회귀 진단

검정

독립성 검정
Dubin-Watson test
-> 통과

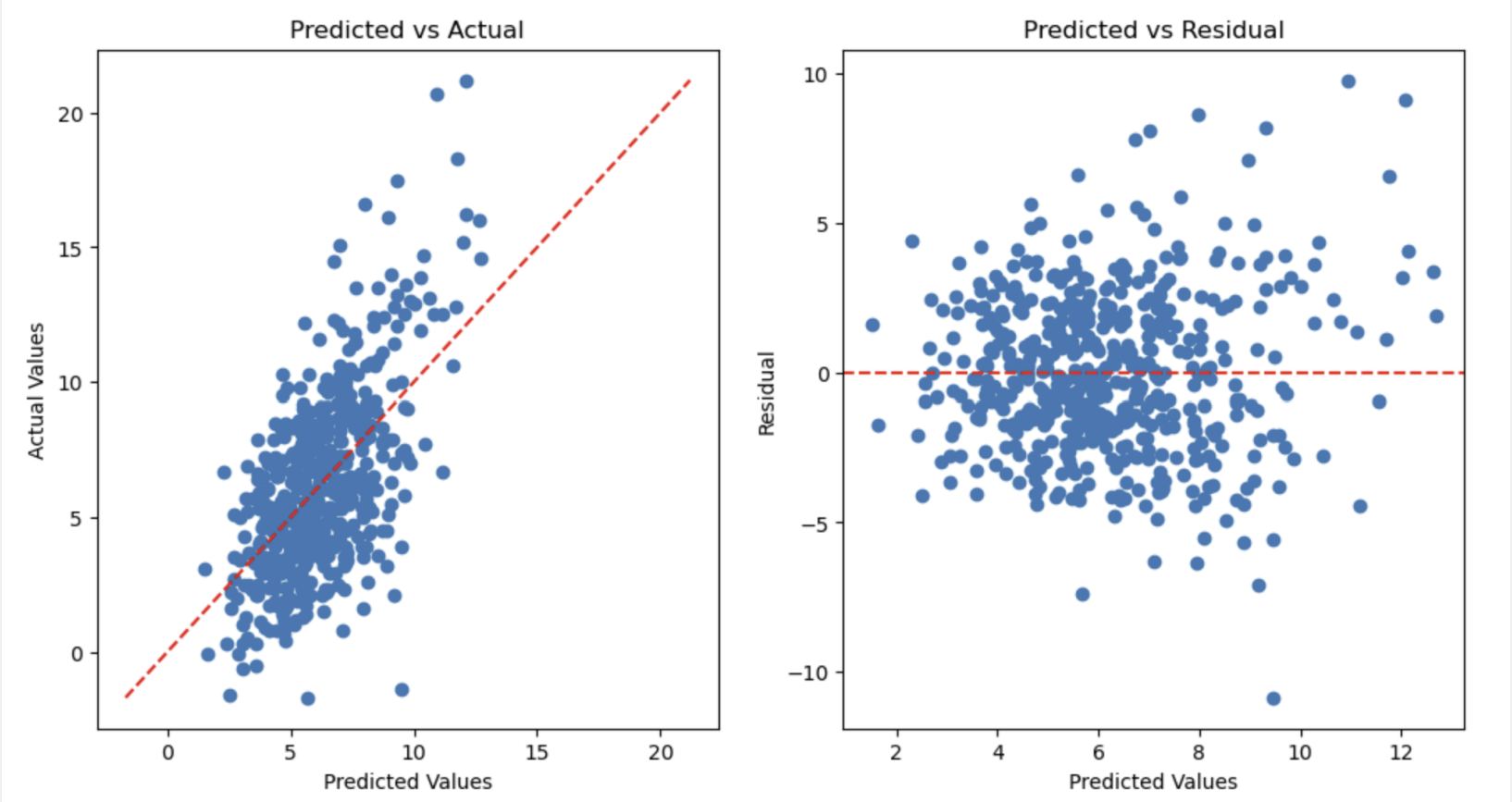
등분산성 검정
Brush-Pagan test
-> 실패

정규성 검정
Kolmogorov-Smirnov test
-> 통과

- 독립성과 정규성은 만족하지만 등분산성은 만족하지 않음
- 감량 구간에 따라 분산이 일정하지 않음을 의미
- 데이터가 일정하지 않고, 특정구간의 감량한 인원이 많아서 발생한 것으로 보임

V 회귀 진단

test



- 처음 분리 해둔 test set으로 테스트 진행
- RMSE: 2.66 -> 평균 오차로 해석
- RMSE 값과 잔차의 표준편차가 크게 차이 나지 않음으로 과적합 된 모델은 아니라고 판단

VI 최종 모델 및 해석

최종모델

회귀식

Y : 초진환자의 3개월 감량

$$Y = 0.8109 - 0.0321 * Age + 0.7430 * YangbangX + 0.0140 * VFA + 0.0911 * Weight + 0.0526 * apedrin1 - 0.0808 * apedrin2 + 0.0133 * Period - 0.0399 * absPeriod + \epsilon \quad , \quad \epsilon \sim N(0,\ 2.57^2)$$

ϵ 의 해석

상위 30% : 1.35kg
상위 10% : 3.29kg -> 노력에 따라 추가 감량 가능성으로 해석 할 만 함
상위 5% : 4.23kg

해석의 유의 사항

- 계수가 체중감량의 영향도를 의미 하지 않음
  - 변수들간의 상관관계가 존재하기 때문
- 등분산성이 만족하지 않아 회귀식에 왜곡이 존재 할 수 있음