2021-2 학기 회귀분석(1139301-01)

기말프로젝트

- 국민건강영양조사 자료를 이용한
청장년층의 운동량에 영향을 미치는 요인에 관한 연구 -

학번 : 20200209
이름 : 이현지

국민건강영양조사는 국민의 건강수준, 행태, 식품 및 영양 섭취 실태에 대한 국가 단위의 대표성과 신뢰성을 갖춘 통계를 산출하고 이를 통해 국민건강증진종합계획의 목표 설정 및 평가, 건강증진 프로그램 개발 등 보건정책의 기초자료로 활용하기 위해 특수 가구(양로원, 군대, 교도소, 외국인 가구)를 제외하고 만 1 세 이상의 모든 가구원을 조사대상자로 선정하였다. 표본 설계 시점에서 가용한 가장 최근 시점의 인구주택총사자료를 기본 추출 틀로 사용하였지만 설계 시점에 따라 표본 추출틀은 유동적으로 바뀌었으며 1998 년에 시행된 제 1 기 조사부터 2019-2021 년 시행된 제 8 기 까지 총 8 기로 이루어져 있다. 조사내용은 가구확인조사, 건강설문조사, 검진조사, 영양조사를 통해 수집된 조사자료로 이루어져있다.

한국표준협회가 발표한 '2018 년 소비 트렌드'에 따르면 전 연령대에서 소비자들이 평소 가장 관심 있는 분야는 '건강(36.6%)'으로 조사됐다. 또한 건강을 1 순위로 꼽은 소비자들이 건강관리를 위해서 하고 있는 활동으로는 '운동(74.8%)'이 압도적으로 많았다. 이처럼 전연령대가 건강관리, 그 중에서도 운동에 관심을 갖고 있는데 그들이 운동에 참여하는 이유는 무엇일까? 운동 참여의 외재적 동기 즉, 심혈관 질환, 정신질환과 같은 질병과 주관적체형인식(체형변화욕구), 거주지역, 성별, 가구소득과 같은 사회경제적 요인들이 어떻게 운동량에 영향을 주는지를 분석하고자 한다.

본 연구는 운동참여에 대한 외재적 동기들을 2019 년 국민건강영양조사에서 6 가지 골라 연령대별 운동량에 영향을 미치는 요인에 관한 연구를 진행 할 것이며 표본은 20 세에서 39 세 사이로 제한하였다. 회귀식은 다음과 같다.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \varepsilon_i$$

변수 설정은 아래와 같이 진행하였다.

운동량을 종속변수로 설정하였고 (중강도 신체활동 일수(BE3_86)) * {60 * 중강도 신체활동 시간(BE3_87)+ 중강도 신체활동 분(BE3_88)} + (고강도 신체활동 일수(BE3_76)) * {60*고강도 신체활동 시간(BE3_77)+ 고강도 신체활동 분(BE3_78)} 이다. 이 중 '모름' 혹은 무응답으로 답한 자는 삭제하였다.

설명변수들은 심혈관질환, 정신질환, 주관적체형인식, 지역, 성별, 가구소득으로 설정하였다. 각각변수들의 설정 방법은 아래와 같다.

심혈관질환은 고혈압 의사진단여부(DI1_dg), 이상지질혈증 의사진단여부(DI2_dg), 뇌졸중 의사진단여부 (DI3_dg), 심근경색증/협심증 의사진단여부(DI4_dg), 당뇨병 의사진단여부 (DE1_dg) 중 하나라도 진단여부가 있음이면 1, 모두가 없음이면 0 으로 설정하였고 모두가 '모름' 혹은 무응답으로 답한 자는 삭제하였다.

정신질환은 우울증 의사진단 여부(DF2_dg) 中 진단여부가 있음이면 1, 없음이면 0 으로 설정하였고 비해당 (청소년, 소아) , '모름'으로 답한자와 무응답자는 삭제하였다.

주관적체형인식은 주관적 체형인식(BO1)의 변수를 사용하였으며 1 에서 5 까지 다섯 척도로 높을수록 비만하다고 인식함을 의미한다. 비해당 (청소년, 소아) , '모름'으로 답한 자와 무응답자는 삭제하였다.

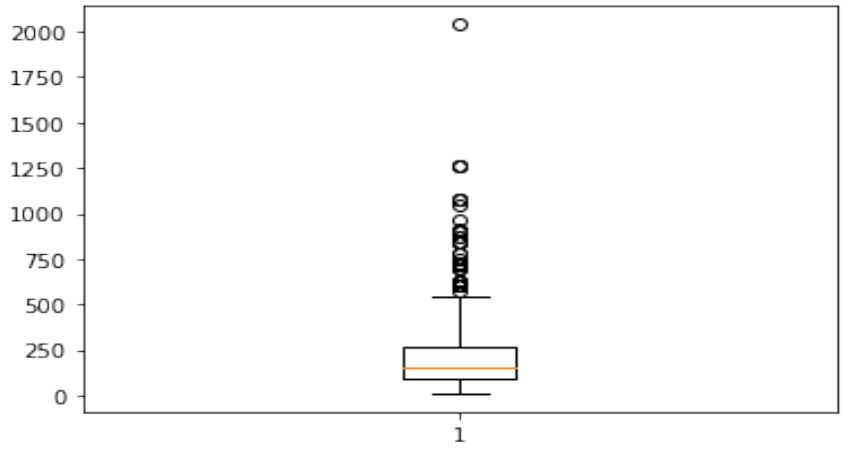
지역은 지역(REGION) 변수를 사용하였고 응답이 서울 혹은 경기이면 '수도권', 울산, 부산, 대구, 인천, 광주, 대전 중 한 곳이면 '광역시', 이외 (강원, 충북, 충남, 전북, 전남, 경북, 경남, 제주, 세종) 중 한 곳이면 '기타지역' 으로 설정하였다. 수도권을 기준으로 광역시, 기타지역 2 개의 지역 더미를 생성하였다.

성별(female)은 성별(SEX) 변수를 사용하였고 1(남성) 이면 0, 2(여성) 이면 1 로 설정하였다. 남성을 기준으로 여성 성별 더미를 생성하였다.

가구소득은 소득 5 분위수(ho_incm5) 변수를 사용하였고 1 에서 5 까지 다섯 척도로 높을수록 고소득자를 의미한다.

종속변수에 무응답자 및 모름이라 답하여 제거한 인원을 제외하고도 남는 결측치(NaN)는 가구소득 변수에서 발견되었는데 5 명 밖에 되지 않아 제거해주었다.

[그림 1] boxplot



IQR 을 이용하여 boxplot 을 그려본 결과이다. 운동량이 극단적으로 높은(주 33 시간 이상) 사람은 이상치로 판단하여 제거해 주었다. (그림 1 참고)

위와 같이 각 변수들에서 삭제해야 할 데이터와 결측치 및 이상치를 제거하고 나면 분석대상자는 전체 585 명이다.

[표 1] 데이터 특성 (명, %)

성별	남성	348(59.48)
	여성	237(40.51)
심혈관질환 여부	있음	24(4.1)
정신질환 여부	있음	16(2.7)
주관적체형인식	매우마른편	14(2.3)
	약간마른편	70(11.9)
	보통	241(41.1)
	약간비만	199(34)
	매우비만	61(10.4)
지역	수도권	306(52.3)
	광역시	132(22.5)
	기타지역	147(25.1)
가구소득	하	26(4.4)
	중하	50(8.5)
	중	116(19.8)
	중상	158(27.0)
	상	235(40.1)

[표 1]에는 데이터의 특성이 나타나 있다. 통계분석에 사용된 총 585 명 중 남성이 59.48%, 여성은 40.51%이다. 심혈관질환 보유자는 4%이고 정신질환 보유자는 2.7%이다. 주관적체형인식은 보통(41.1%)이 많고 다음으론 약간비만(34%)가 많았다. 지역은 수도권 52.3%, 광역시가 22.5%, 그 외지역이 25.1%이다. 가구소득 5 분위에서는 하위가 4.4%에 불과하고 소득이 높아질수록 점점 많아져 상위가 40% 달하여 골고루 분포하지는 않는다.

[표 2]평균운동량 (시간/주)

전체	3.40	
성별	남성	3.70
	여성	2.97
심혈관질환 여부	있음	3.54
	없음	3.40
정신질환 여부	있음	2.26
	없음	3.44
주관적체형인식	매우마른편	3.05
	약간마른편	3.07
	보통	3.49
	약간비만	3.48
	매우비만	3.26
지역	수도권	3.09
	광역시	3.61
	기타지역	3.90
가구소득	하	4.17
	중하	3.56
	중	3
	중상	3.36
	상	3.52

[표 2]에서는 주당 평균운동시간을 보았다. 585 명 전체 평균 주당 운동시간은 3.4 시간인데 남성은 주당 평균 3.7 시간으로 여성보다 0.73 시간이 많았다. 심혈관질환이 있는 사람(평균 3.54 시간)은 그렇지 않은 사람들(3.40 시간) 보다 평균 0.14 시간이 많았다. 한편, 정신질환(우울증)이 있는 사람은 오히려 정신질환이 없는 사람보다 운동시간이 적었는데 이는 의사의 권유에도 불구하고 운동을 하기 쉽지 않은 심리상태를 반영한 것 일수도 있다. 그리고 앞에서 본 바와 같이 정신질환자는 총 16 명에 불과하다. 주관적체형인식에 따른 운동량의 차이에서도 비만의 인식이 높을수록 운동량이 많아지는 선형의 관계는 아닌 것으로 나타났다. 다만 자신이 매우 마른편이라고 생각하는 사람들의 평균운동시간은 3.05 시간으로 전체평균 3.40 시간보다 현저히 낮게 나타났다. 지역별로는 수도권이 낮게 나타났고(3.09 시간) 광역시(3.61 시간)와 기타 지역(3.90 은)로 수도권의 평균운동시간이 적게 나타났다. 가구소득분위와 운동시간과의 관계에서도 선형적인 관계를 발견하기 어렵다. 소득 1 분위(하위 20%, 4.17 시간))와 5 분위(상위 20%, 3.56 시간)가 가장 높게 나타났다.

[표 3] Spearman 상관분석

	운동량	심혈관 질환	정신질환	주관적 체형인식	광역시	기타지역	여성	가구소득
운동량	1	0.04	-0.08	0.02	0.04	0.07	-0.16	0.02
심혈관질환	-	1	0.02	0.18	-0.02	-0.01	-0.05	0.05
정신질환	-	-	1	0.01	-0.07	0.04	0.05	0
주관적 체형인식	-	-	-	1	0.03	0	-0.02	0.04
광역시	-	-	-	-	1	-0.31	-0.02	-0.11
기타지역						1	-0.01	-0.03
Female(여성)	-	-	-	-	-		1	0.02
가구소득	-	-	-	-	-		-	1

회귀분석을 실시하기 전 변수들 간의 상관분석을 실시하여 그 결과가 [표 3]에 나타나 있다. 우선 모든 독립변수들이 명목척도(더미변수)이거나 서열척도 변수이어서 Pearson 상관관계는 별 의미가 없으므로 Spearman 상관관계를 보았다. 광역시와 기타지역 간의 상관계수 -0.31 을 제외하면 매우 낮은 상관관계를 보여 다중공선성(multi-collinearity)의 문제는 없는 것으로 나타났다.

[표 4] 회귀분석 결과

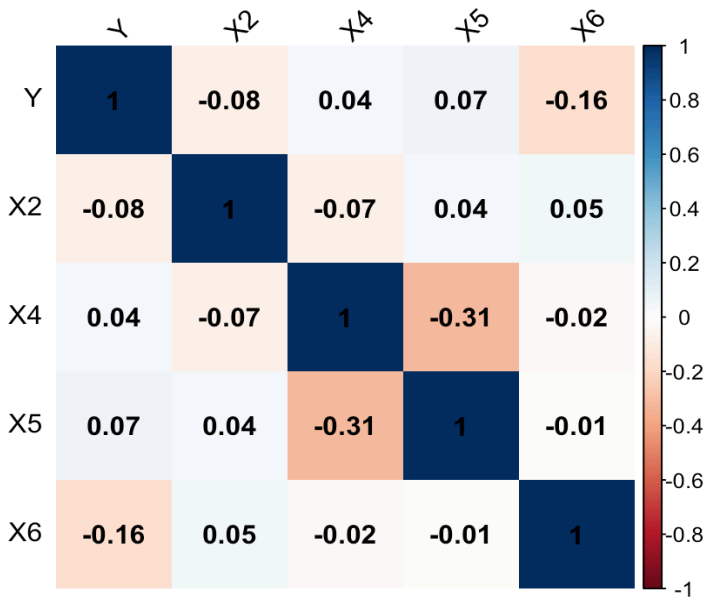
	전체 (n=585)
intercept	3.09796 (4.87e-07 ***)
심혈관질환 (X1)	0.05734 (0.92245)
정신질환 (X2)	-1.04456 (0.13919)
주관적체형인식 (X3)	0.05409 (0.67303)
광역시 (X4)	0.48464 (0.08684)

기타지역 (X5)	0.79795 (0.00594 **)
Female(여성) (X6)	-0.60887 (0.00279 **)
가구소득 (X7)	-0.70310 (0.72490)
결정계수	0.03505
수정결정계수	0.02334

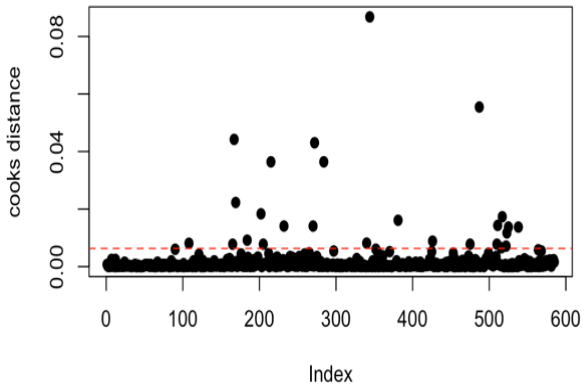
다음으로 회귀분석을 실시하였다. 3% 정도에 불과한 결정계수로 알 수 있듯이 운동량에 대한 독립변수들의 설명력은 전반적으로 낮은 것으로 나타났다. 변수 별로 보면 기대와는 달리 심혈관질환여부, 정신질환여부 뿐만 아니라 주관적체형인식 조차도 통계적으로 의미있는 설명변수가 되지 못하고 있다. 앞 [표 2]의 설명에서 보았듯이 주관적체형인식을 매우 마른편과 나머지로 분류하였더라면 적어도 유의미한 결과가 나왔을 지도 모르겠지만 그것은 너무 자의적인 구분이 될 것이다. 수도권 사람들에 비해서 광역시와 기타지역에 거주하는 사람들의 운동량이 통계적으로 의미있게 높게 나왔으나 이에 대한 마땅한 설명도 어렵다. 성별로는 [표 2]에서와 마찬가지로 여러 다른 변수들을 통제한 다중회귀분석에서도 여성의 운동량이 적은 것으로 나타났다. 결론적으로 지역과 성별 이외에는 통계적으로 의미있는 독립변수는 없어 심혈관질환, 정신질환의 여부나 자신의 체형에 대한 인식조차도 운동량에 통계적으로 유의미한 영향을 주지 못하는 것으로 나타났다.

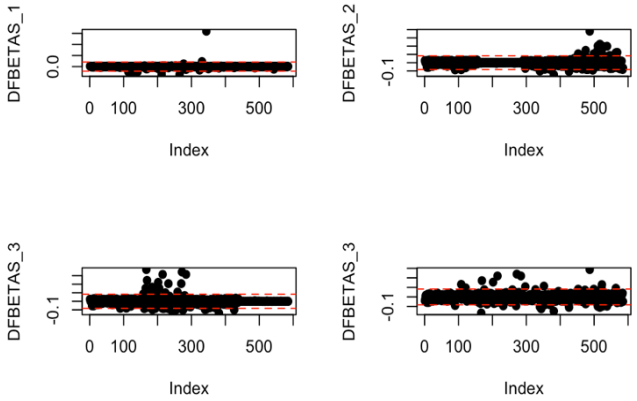
유의하지 않은 변수가 많았기 때문에 forward, backward, stepwise 세 가지 방법의, 변수 선택방법으로 변수를 선택해 보았고 그 결과 수정된 결정 계수가 가장 높은 모형을 기준으로 정신질환, 광역시, 기타지역, female 이 변수로 선택되었다. 다음으론 새롭게 정의된 모형이 회귀분석의 가정을 잘 따르는지 회귀 진단을 진행하였다. 먼저 다중공선성에 대해 알아보았고 광역시와 기타지역 간의 상관계수 -0.31 을 제외하면 매우 낮은 상관관계를 보여 다중공선성의 문제는 없는 것으로 나타났다. (표 4 참고)

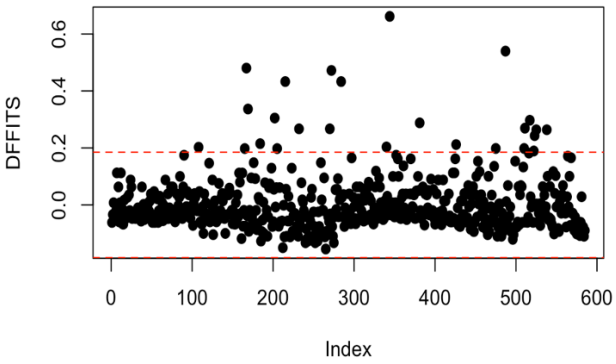
[표 4] Spearman 상관분석 (새롭게 정의된 모형)

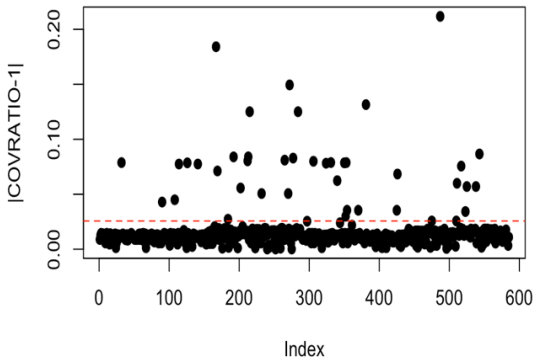


다음으론 영향력 관측치를 살펴 보았다. 영향력 측도론 잔차와 leverage 를 동시에 고려한 척도인 Cooks distance(쿡의 거리), βj hat 에 대한 영향력을 측정 하는 DFBETAS, Yi hat 에 대한 영향력을 측정 하는 DFFITS, $Cov(\beta$ hat)의 추정값에 대한 i번째 관측치의 영향력을 측정하는 COVRATIO 를 사용하여 분석을 진행 하였다. 아래의 그래프들은 순서대로 Cooks distance, DFBETAS, DFFITS, COVRATIO 이다.







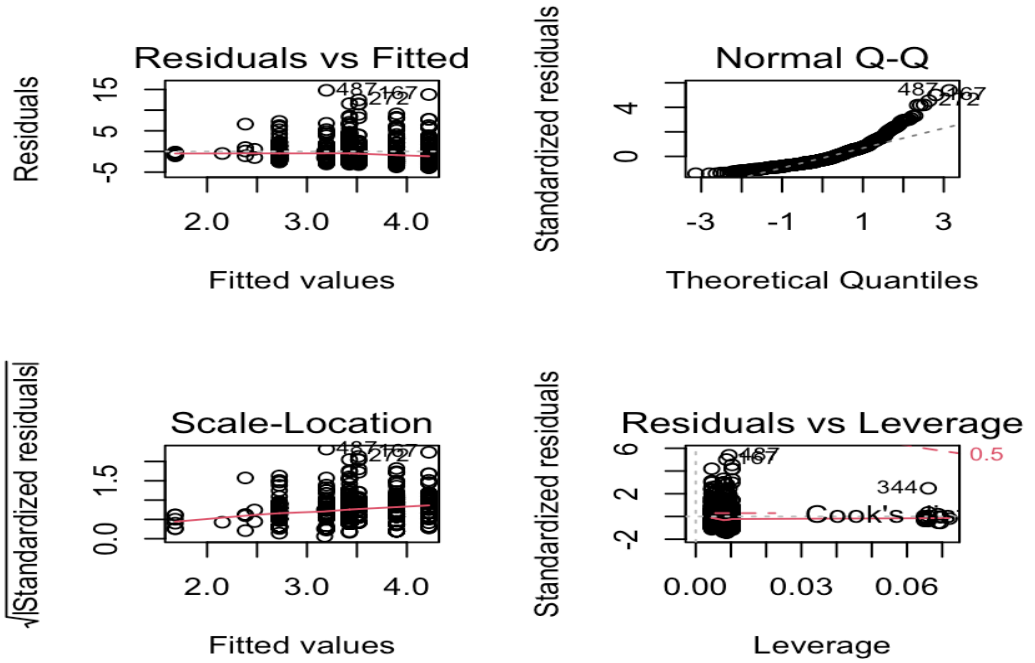


그래프를 분석해보면, 기준이 되는 붉은 선 밖에 있는 관측치들이 위 그래프에서 보여지는 각 기준에 따라 진단을 진행 하였을 때의 이상치이다. 전체적으로 이상치로 판단되어 지는 값들이 많아 모두 제거하는 것을 고려해 보았으나 이는 데이터를 왜곡하는 것으로 판단되어져 눈에 띄는 관측치들인 487,272,167,344,381,517,169,426,202,511,525,232,538,340,270,284,215,108,385,523 을 지우고 총 565 개의 관측치를 다시 회귀 모형에 적합을 시켜보았다. 이상치 제거 후 수정된 결정계수는 0.03843 으로 회귀 진단 전 회귀식 보다 개선되었음을 확인하였다. ([표 5] 참고)

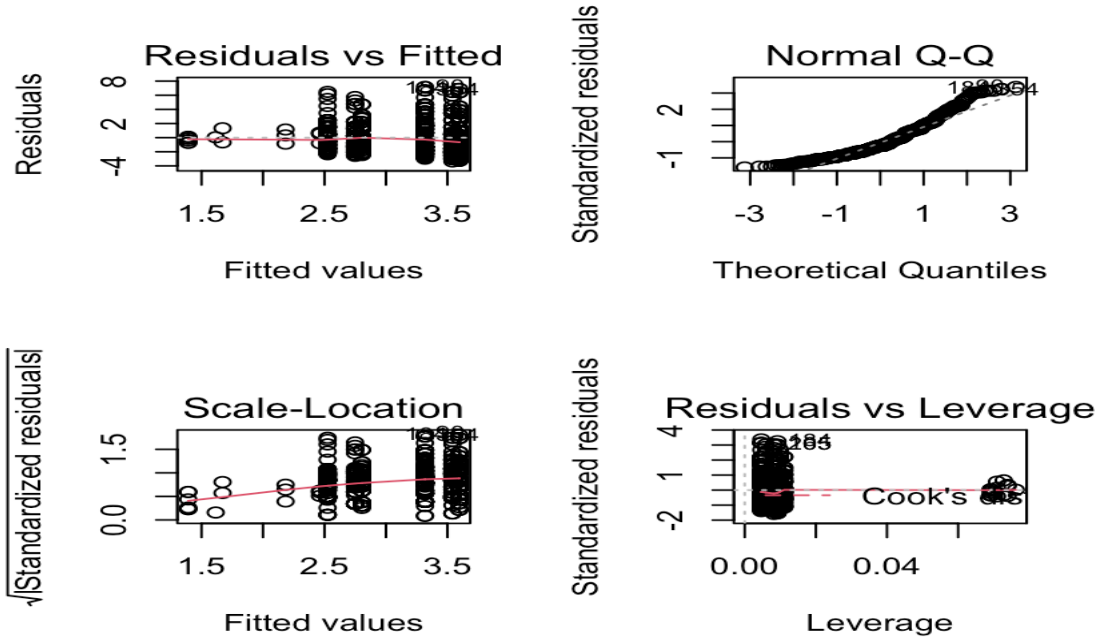
[표 5] 수정된 모형의 회귀분석 결과

	전체 (n=565)
intercept	3.3171
정신질환 (X2)	-1.133 (0.0437 *)
광역시 (X4)	0.2794 (0.2206)
기타지역 (X5)	0.2238 (0.3053)
Female(여성) (X6)	-0.7919 (1.85e-05 ***)
결정계수	0.04525
수정결정계수	0.03843

이상치 제거 전



이상치 제거 후



이상치 제거 후를 비교보면 등분산성, 정규성, 선형성 면에서 미흡한 점이 있지만 이상치 제거 전과 비교해 보았을 때 상당히 개선된 부분이 보여 해당 모형을 최종 모형으로 결정하였다.
다음으론 최종 선택된 모형에 대한 타당성을 검증하는 과정을 거쳤다.

훈련자료(70%)와 확인자료(30%)로 나누어 예측 오차를 구하였다. $PRESS$ = 2583, SSE = 2545 로 SSE 와 PRESS 가 거의 비슷한 것을 확인하여 예측의 정확도가 높으며 $R2\ Predic$ = 0.031, R2 = 0.045 로 PRESS 를 이용하여 구한 R2 값과 예측 값을 구하지 않은 R2 이 서로 비슷한 것이 확인되어 설명력이 급격히 떨어지진 않았고 모형이 예측 모형의 확인 면에서도 좋다는 것을 확인하였다. 따라서 수정된 모형을 최종 회귀식으로 결정하였다. 최종 선택된 회귀식은 아래와 같다.

$$운동량 = 3.31 - 1.1\ 정신질환 + 0.27\ 광역시 + 0.22\ 기타지역 - 0.79 female$$

회귀 진단을 거쳤음에도 불구하고 3.8% 정도의 결정계수로 알 수 있듯이 운동량에 대한 독립변수들의 설명력은 전반적으로 낮은 것으로 나타났다. [표 5]의 회귀분석을 살펴보면, 정신질환이 있는 사람이 없는 사람보다 약 -1.1 시간 정도 운동량이 적고 광역시 거주인은 수도권 거주인 보다 약 0.3 시간 , 기타지역 거주인은 수도권 거주인 보다 약 0.22 시간 운동량이 많다는 것을 알 수 있다. 그러나 P value 를 보면 광역시와 기타지역은 설명력이 낮은 변수임을 알 수 있다. 그나마 의미 있는 변수는 female(성별) 인 것으로 나타났는데, 이는 여성이 남성보다 약 -0.79 시간 정도 운동량이 적은 것을 나타낸다.

이 연구를 진행하면서 회귀식이 가정에 위배되어 어려움이 있었다. 회귀 진단을 하면서도 이상치로 판단되는 관측치가 너무 많았기 때문에 그것들을 모두 삭제하고 진행해야 하나 고민하였지만, 이상치를 많이 삭제하는 것이 데이터를 왜곡하는 것이라는 판단 하에 눈에 띄는 관측치만 삭제를 진행하였다. 또한 더빗왓슨검정도 진행하였는데 더빈 왓슨 통계량은 1.88 로 2 와 가까이 나타났지만 p-value 가 0.0695 로 잔차 끼리 자기상관성이 있다고 검정 되어 어려움이 있었다. 이 주제를 선택한 이유는 청장년층 중 특히 여성들이 수치적인 몸무게와는 무관하게 주관적인 체형 인식 하에 자신은 체중 조절이 필요하다고 판단하여 운동량이 높을 것이라는 회귀분석 결과를 도출하여 주관적 체형 인식이 유의미한 변수인지 확인해 보고 싶어서 였다. 기대와는 다른 회귀분석 결과가 나왔지만 추후 이를 보정하여 더 전문적인 분석 방법을 통해 다시 한번 진행해 보고 싶은 마음이 있다.