프로젝트 보고서

비타민 D 의 농도와 혈당의 연관성

이현아 | 2018.06.11

프로젝트 목적

평일에 통학과 아르바이트 때문에 어쩔 수 없이 대중교통속에서 사람들에게 이리저리 치이다 보니 피로에 찌들어 주말과 공휴일에는 집에만 있는 '집순이'가 되곤 했다. 그랬더니 가족부터 친구들까지 "그러다가 우울증 걸린다." 라는 이야기를 꽤나 많이 듣게 되었다. 대충 각종 매체를 통해 햇볕속의 '비타민 D' 때문이라는 것은 알았지만 도대체 비타민 D 와 우울증이 무슨 관계가 있는지 궁금해서 찾아보았고 그 내용은 다음과 같았다.

"일조량이 감소하면 그에 따라 햇볕을 통해 흡수할 수 있는 '비타민 D'의 결핍이 일어난다. 비타민 D 는 항우울 작용을 하는 도파민과 세로토닌의 합성에 관여하는데 비타민 D 가 부족하여 합성이 제대로 이루어지지 않아 우울증이 발병하는 것이다. 우울증 뿐만 아니라 비타민 D 의 결핍은 다양한 질병으로 이어질 수 있다."

이때 나의 이목을 끌었던 부분은 바로 조사의 마지막 줄이다. 그저 우울증만 관련이 있는 줄 알았는데 다른 다양한 질병들과 관련이 있다니. 특히 비타민 D 의 결핍이 당뇨에도 영향을 준다는 점에서 놀랐다.

"국민 건강 영양 조사"를 살펴보니 혈중 비타민 D 의 농도와 당뇨병 여부의 근거가 되는 혈당을 포함하고 있었고, 혈중 비타민 D 의 농도와 혈당의 연관성을 살펴보는 것을 프로젝트의 주제로 선정하게 되었다.

혈중 비타민 D 의 농도가 낮으면 인슐린의 분비가 줄어들고 혈당을 낮추는 인슐린의 기능이 떨어져 세포가 포도당을 효과적으로 연소하지 못하게 된다. 따라서 당뇨가 발병할 가능성이 높아진다는 과학적 사실을 바탕으로 실제로 비타민 D 의 농도가 높으면 혈당이 줄어드는지 확인하고자 한다. 따라서 프로젝트의 가설은 다음과 같다.

- 혈중 비타민 D 의 농도가 높을수록 혈당이 낮을 것이다
- 독립변수 (x) : 혈중 비타민 D 의 농도 (ng/mL)
- 종속변수 (y) : 공복혈당 (ng/dL)

프로젝트 분석

R 프로그램을 이용하여 데이터를 분석하여 가설을 확인하고자 한다.

먼저 데이터 'knhanes.csv'를 read.csv 함수를 이용하여 불러왔다. 이때 EOF error 가 발생했고 이는 자료안에 " "를 포함한 데이터값이 존재한다는 것을 의미하므로 함수에 quote=""를 추가적으로 적어주고 에러를 해결하였다.

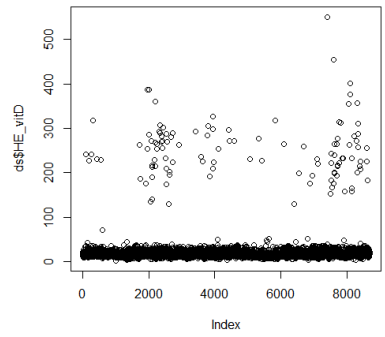
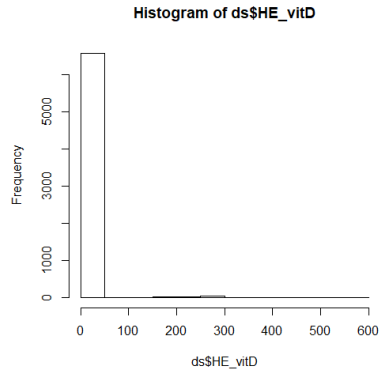
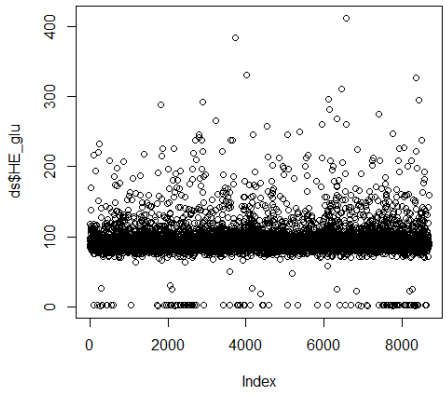
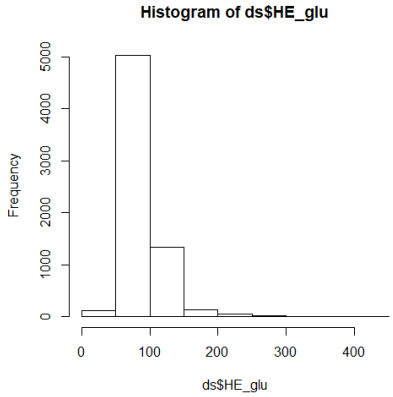
분석에 앞서 독립변수 비타민 D (HE_vitD)와 종속변수 혈당 (HE_glu)에 is.na 함수를 적용시키고 table 함수에 적용하여 결측(NA)을 제외한 데이터 개수를 구했다.

	데이터 개수 (개)
HE_vitD	6704
HE_glu	6698

다음으로 summary 함수를 이용하여 평균, 중앙값, 최댓값, 최솟값을 sd 함수를 이용하여 표준편차를 구하여 기초 통계량을 파악하였다. 이때 na.rm=T 를 이용하여 결측(NA)을 제거하였다.

	평균	표준편차	중앙값	최댓값	최솟값
HE_vitD	21.32	31.94	16.82	1.26	551
HE_glu	95.04	25.29	91	1	412

데이터의 분포를 그림으로 확인하기 위해 plot 과 hist 함수를 이용하여 두 변수의 분포를 확인했다.

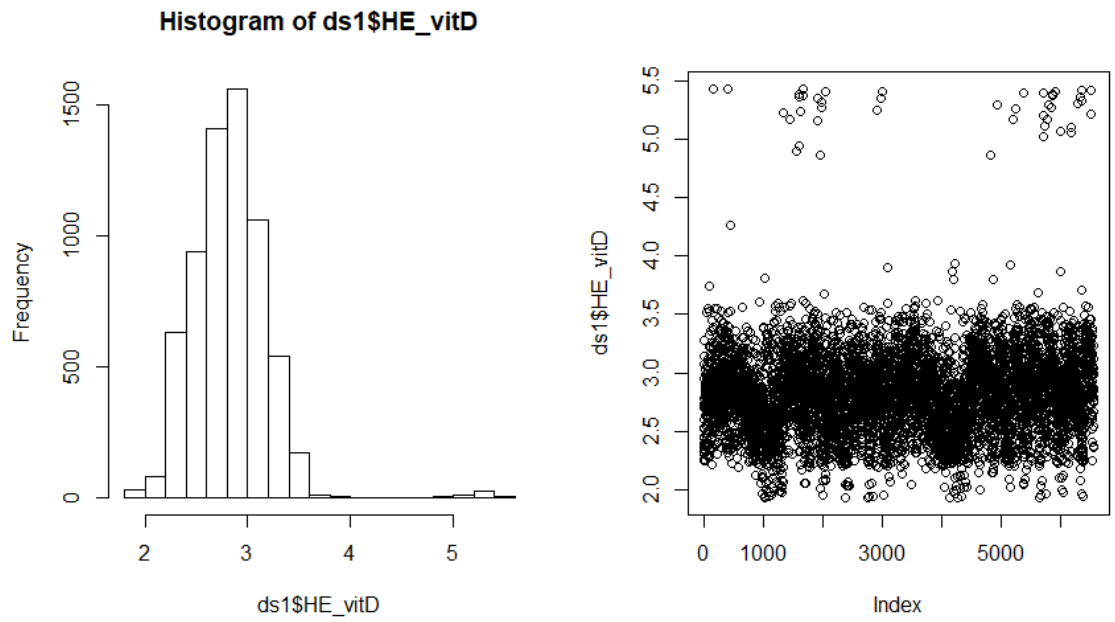
	Plot	hist
HE_vitD		
HE_glu		

이때 이상치 제거를 위해 quantile 함수에 probs=c(1,99)/100 을 입력하여 상위, 하위 1% 값을 구하고, ifelse 함수를 이용하여 제거하였다. 이때 값을 제거하면서 데이터의 개수는 다음과 같이 바뀌었다.

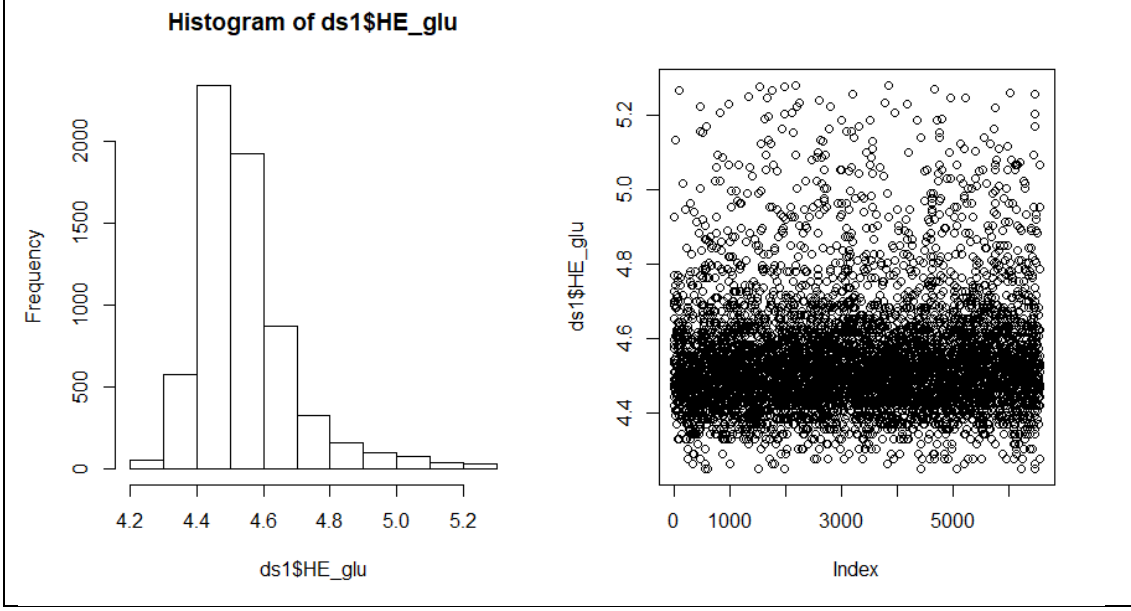
	데이터 개수
HE_vitD	6496
HE_glu	6496

또한 이상치를 제거한 그래프는 다음과 같으며 필요에 따라 log 함수를 이용하여 로그 치환을 실행했다.

HE_vitD

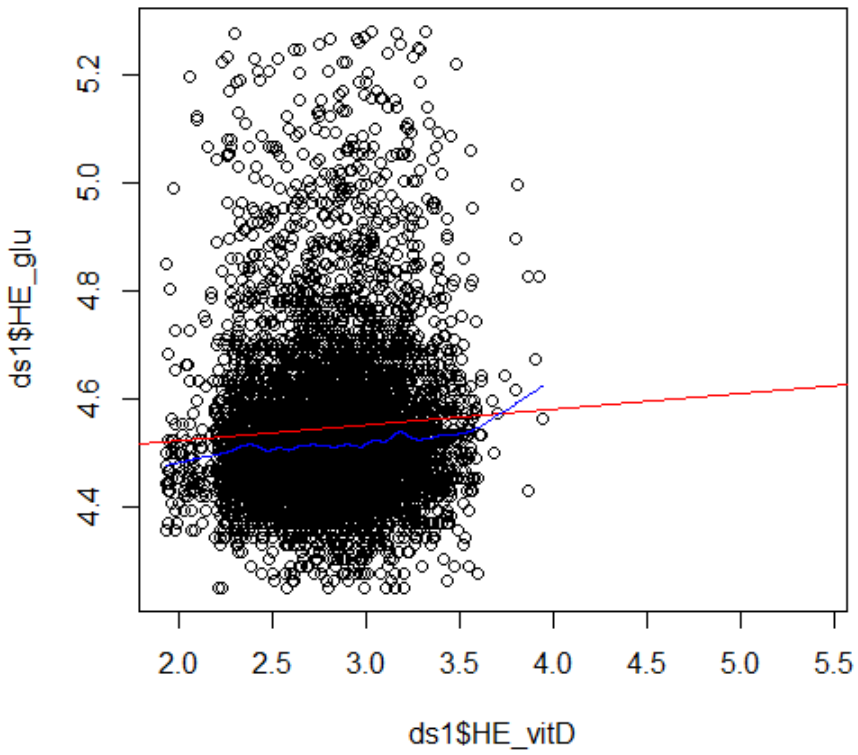


HE_glu



데이터 처리를 모두 마쳤으므로 plot, abline 함수를 이용하여 두 변수의 연관성을 그래프로 나타냈다.

lowess 함수를 이용하여 line 을 그리기 전에는 complete.cases 를 이용하여 데이터를 선택하여 그래프를 그렸다.



위의 그림에서 그려진 abline 을 통해 비타민 D 와 혈당이 양의 관계를 갖음을 알 수 있다. 즉, 혈중 비타민 D 의 농도가 높을수록 혈당이 높다는 의미인데 이는 가설에 반대이다. 따라서 이 분석이 신뢰도와 설명력이 어느정도인지 확인하기 위해 선형회귀모형을 분석했다.

선형 회귀 모형은 lm(y~x, data=ds) 함수를 이용하였고 비타민 D 와 혈당의 회귀 모형을 m1 이라고 지정하고 summary 함수에 대입하였다.

유의수준을 0.05 라고 가정했을 때, P-value 는 8.586*(e^(-7)) = 0.00782941856 로 0.05 보다 작은 값을 갖고, R-squared 값은 0.003756 으로 대략 0.004 즉 0.4 퍼센트의 설명력을 가진다. 따라서 양의 관계를 이룬다는 분석은 크게 설명력이 없음을 알 수 있다.

마지막으로 다변량 분석을 해보았다. 프로젝트에서 종속변수에 해당하는 변수가 공복혈당이므로 이와 관련이 있는 독립변수들을 찾아야 했다. 어떻게 할지 고민하다가 왠지 콜레스테롤이 높으면

혈당이 높지 않을까 혹은 체중이 많이 나가면 혈당이 높지 않을까 하는 추측을 기반으로 선정했다.

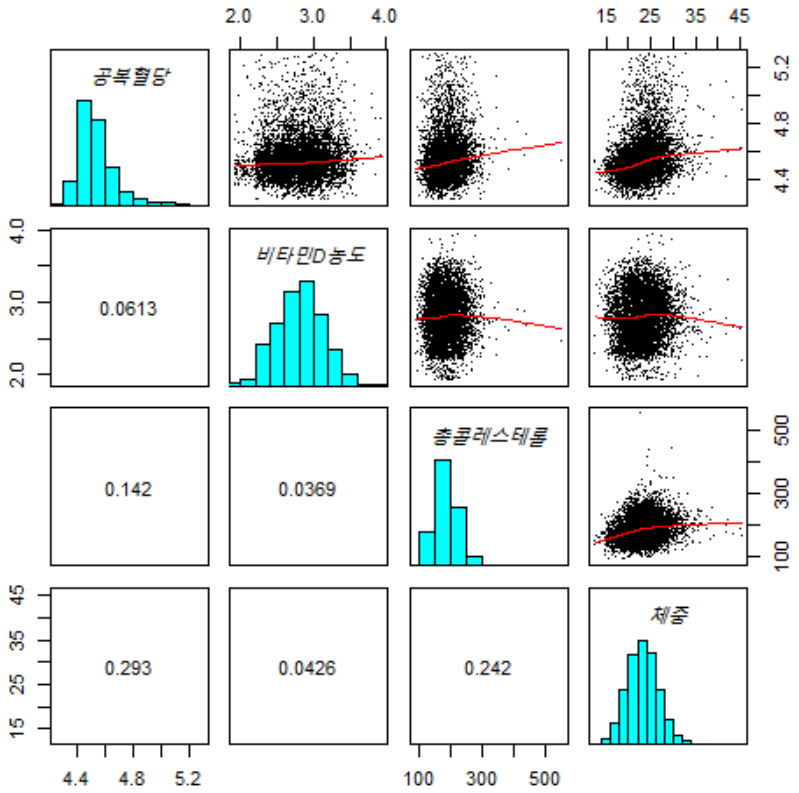
먼저 추측에 근거가 있고, 어느정도의 설명력이 있는지 확인하기 위해 선형 회귀 모형을 각 변수마다 설정하고 분석했다.

첫번째, lm 함수를 이용하여 m2 라고 지정한 콜레스테롤과 혈당의 선형 회귀 모형의 분석은 다음과 같다. 비타민 D 와 같이 유의확률을 0.05 라고 가정했을 때, P-value 는 2.2*(e^(-16)) = 0.00000024758 로 0.05 보다 매우 작았다. R-squared 값은 0.02016 대략 2%의 설명력을 가진다.

두번째, lm 함수를 이용하여 m3 라고 지정한 체중과 혈당의 선형 회귀 모형의 분석은 다음과 같다. P-value 는 콜레스테롤의 경우와 같았고, R-squared 값은 0.08585 대략 8 퍼센트의 설명력을 가진다.

마지막으로 lm 함수를 이용하여 m4 라고 지정한 60 초 맥박수와 혈당의 선형 회귀 모형은 다음과 같다. P-value 는 0.4207 로 0.05 보다 크기 때문에 귀무 가설을 기각할 수 없고, R-squared 값은 0.00114 대략 0.1%의 설명력을 가진다.

따라서 귀무 가설을 기각할 수 없는 변수 60 초 맥박수를 제외한 나머지 변수를 pairs 함수로 묶어준 뒤 plot 함수를 이용하여 그래프를 그렸다.



분석 결과

이 프로젝트의 가설인 '혈중 비타민 D 의 농도가 높을수록 혈당이 낮을 것이다.' 선형 회귀 분석 결과 비타민 D 의 농도가 높을수록 혈당 또한 높다는 결과가 도출되었지만 이는 설명력이 크게 높지 않았다. 또한, 비타민 D 로 혈당을 설명하기엔 부족하다는 것을 알 수 있다.

따라서 비타민 D 의 농도로 혈당을 설명하기엔 적절하지 않으므로 비타민 D 와 혈당은 직접적인 관계가 크게 없음, 혈당에 대한 비타민 D 의 영향력이 크지 않음을 알 수 있다.

다변량 분석을 통해 체중이 비타민 D 보다 혈당에 대한 설명력과 영향력이 높다는 것을 알 수 있었다.

자료 출처

https://terms.naver.com/entry.nhn?docId=3409186&cid=58413&categoryId=58413

http://news20.busan.com/controller/newsController.jsp?newsId=201805300000188