**2021-2 회귀분석(1139301-01)**

**Final project**

**제출기한: 12월 19일까지**

**기말 프로젝트 작성 방법**

1. 미리 작성한 연구계획서를 바탕으로 서론, 본론, 결론의 구성으로 작성합니다.

2. 자료는 국민건강영양조사 제8기 1차년도(2019년)자료를 이용합니다. (모름, 무응답은 결측치로 간주하고, 결측치는 모두 제거하고 분석합니다.)

자료명: Hn19\_all.csv

3. 글자포인트는 10pt로 설정하고, 그림을 포함하여 10페이지 이내로 작성합니다.

4. 통계프로그램으로 결과를 산출시, 가능하면 패키지와 함수를 이용하여 결과를 산출합니다.

5. 본 프로젝트에 이용된 통계프로그램의 코드도 함께 제출하시기 바랍니다.

**I 서론**

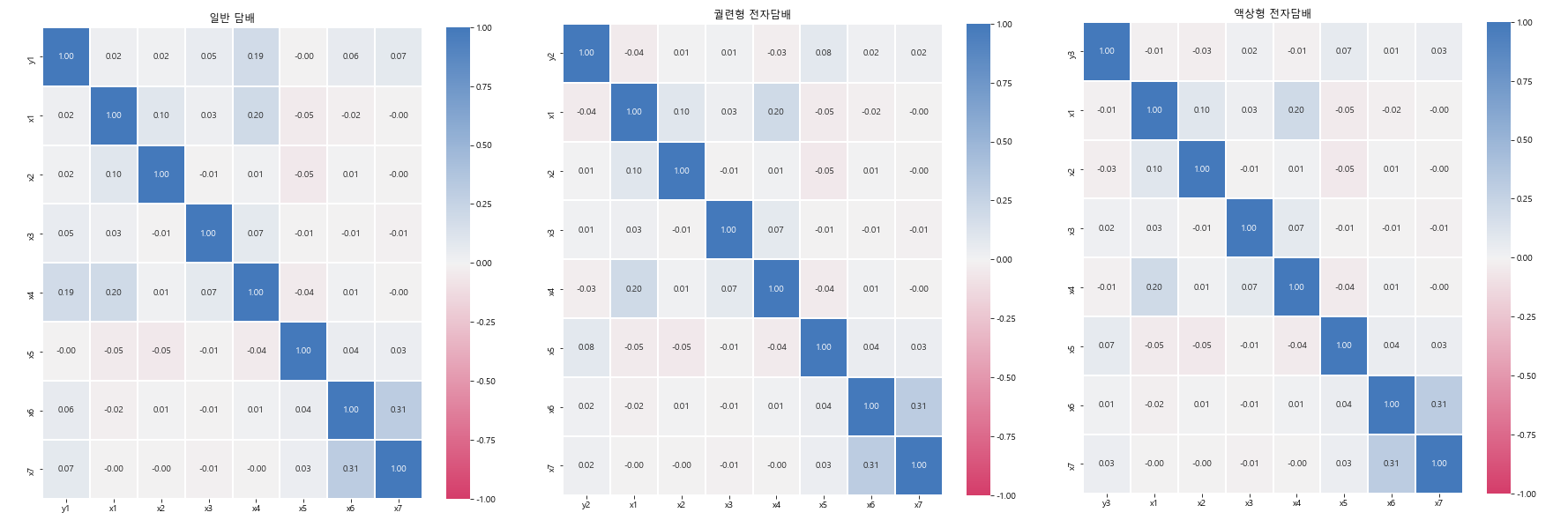
국민건강영양조사는 국가단위의 대표성과 신뢰성을 갖춘 통계를 산출함으로써 국민건강증진 종합계획의 목표 설정 및 평가를 위한 기초자료로서 매년 시행되고 있는 조사이다. 되도록 국민의 대표성이 있는 표본을 추출하였으며 자체적인 설문조사, 혈압검사, 혈액검사 등 총 13가지의 자료를 수집하고 있다. 이는 가구원확인조사에서도 추출률/응답률 계산 및 가중치 산출 등을 위해서도 이용된다. 조사에서도 추가되었듯 국민들의 건강에 대한 관심이 증가함에 따라 기호식품인 담배의 소비가 줄어드는 추세이며 건강을 위해 연초보다 현저히 낮은 독성물질을 포함하고 있는 전자담배가 대체품으로써 등장하였다. 본 연구에서는 국민건강영양조사에 포함된 담배의 종류에 따른 흡연유무와 국민들의 건강데이터를 활용해 담배의 종류별 건강에 끼치는 악영향에 대한 상관관계를 “R”과 “파이썬” 프로그램을 이용해 분석해보고자 한다.

**II 본론**

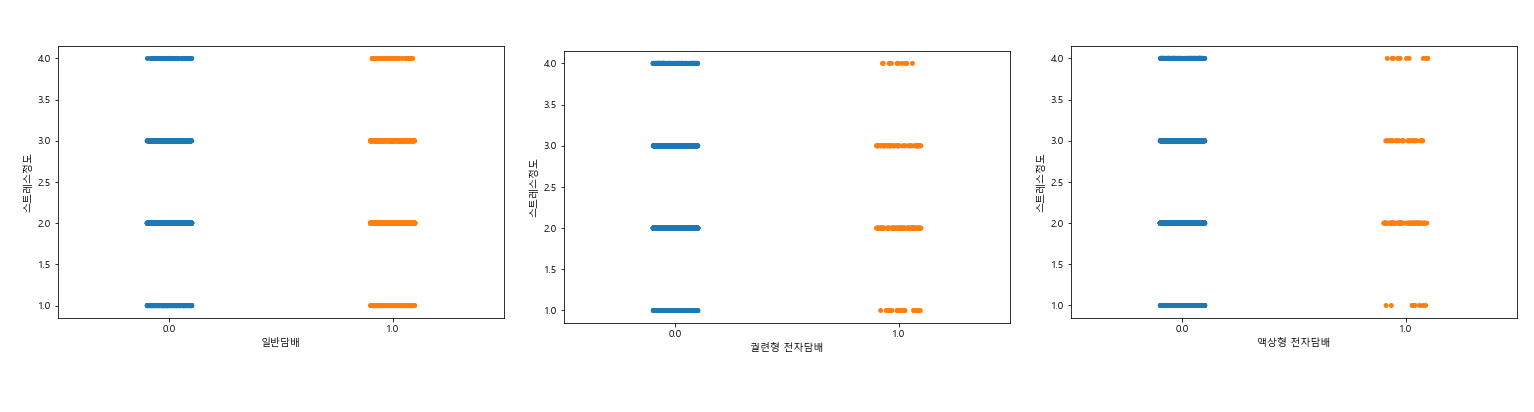
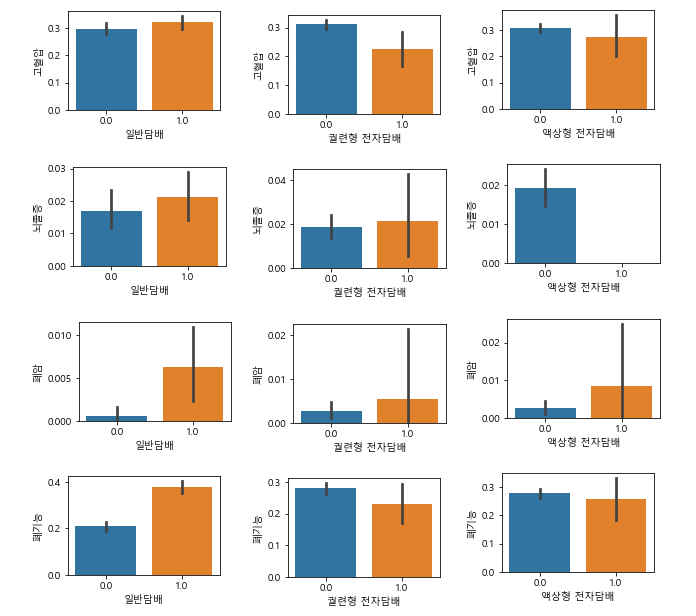
1. 데이터 전처리

데이터는 국민건강영양조사에서 제공한 “Hn19\_all.csv”파일을 이용하며 데이터 전처리에는 파이썬을 주로 사용하였다. 데이터 내의 담배 중 3가지 종류인 “일반 담배”(BS1\_1), “궐련형 전자담배”(BS12\_37), “액상형 전자담배”(BS12\_1)를 종속변수로 사용하며 주로 흡연과 인과관계가 크다고 알려진 “고혈압”(DI1\_dg), “뇌졸증”(DI3\_dg), “폐암”(DC6\_dg), “폐기능이상”(HE\_COPD), “스트레스정도”(BP1)를 독립변수로 사용하며 필요시 “신체활동여부” 데이터인 BE3\_71과 BE3\_81 또한 확인한다. 해당데이터만을 추출 후 새로운 데이터프레임을 만들어 전처리시 기타 미사용 변수들에 영향을 끼치지 않도록 하였다. 본래 연구계획서에 포함하였던 BS3\_1, BS12-47, BS12\_2는 너무 과다한 결측치 및 설문내용의 모호함으로 인해 사용하지 않는다. 분석 시 편의를 위해 종속변수는 순서대로 y1, y2, y2의 변수명을, 종속변수는 x1, x2, x3, x4, x5, x6, x7의 변수명을 붙여 사용하였으며 질병유무에 따라 병이 있다면 1의 값을, 병이 없다면 0의 값을 갖도록 처리하였다. 설문내용의 답을 모르거나 성인이 아닌 경우 결측치와 같이 처리하고 제거하였으며 “스트레스정도”의 경우 가장 낮은 경우를 1, 최대 4까지 정수형으로 올라가도록 값을 조정하였다. 변수의 변환 전, 궐련형과 액상형 전자담배에서 100~200명대의 인원을 가진 것과 달리 데이터의 전처리를 마친 후 담배종류당 각 3170명으로 증가하여 해당 데이터로 실험을 진행하였다.

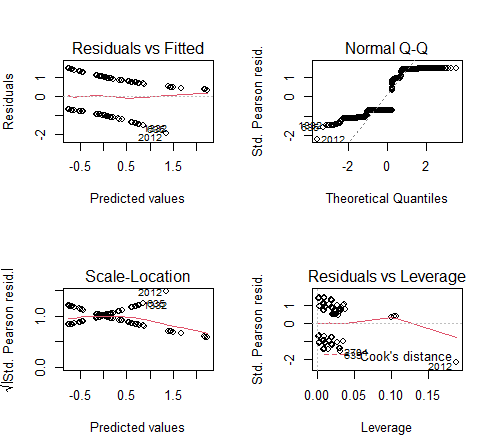
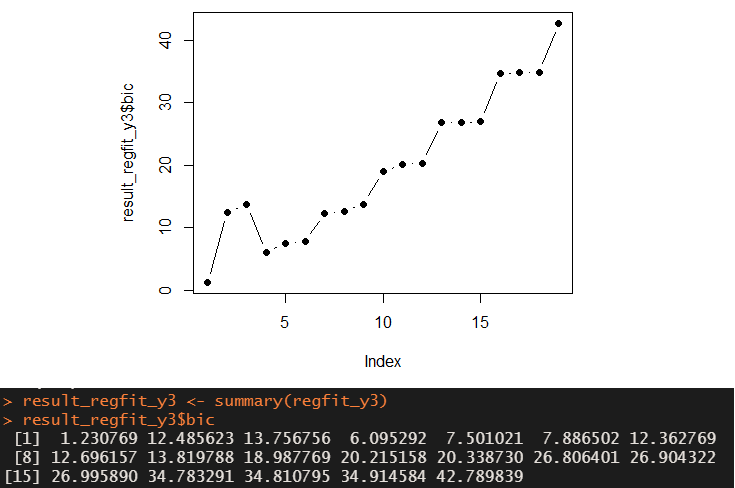
1. EDA

파이썬에서 실행한 Heatmap 검사에 따르면 변수들간의 상관관계는 미미하다. 모든 담배에서 “고혈압”과 “폐기능”의 관계가 가장 크다고 나오나 0.2의 수준에 그쳐 유효한 값은 아니라 판단된다.

담배 별 각 질병의 환자를 전부 확인해 본 결과 일반담배의 경우 예상대로 모든 질병에서 흡연자가 많이 분포되어 있는 것이 확인되며 궐련형 전자담배는 “고혈압”과 “폐기능” 질병에서 오히려 비흡연자보다 적은 인원이 포함되었으며 전체적으로 환자의 비율이 일반담배보다 줄어든 형태를 보여준다. 액상형 전자담배는 “뇌졸증” 질병에서 특히 눈에 띄는 변화를 보인다. 그나마 상관관계가 높았던 “고혈압”과 “폐기능”이 “뇌졸증”과 “폐암”보다 흔하다는 점을 감안해 볼 때 중증질병으로 취급되는 “뇌졸증”과 “폐암”의 감소는 일반 담배와 전자담배의 실질적인 차이라 보여지기는 하나 “폐암”의 흡연자/비흡연자 비율은 여전히 전자담배가 무해하지 않다는 점을 보여준다.

개인이 판단하는 “스트레스정도”는 흥미롭게도 일반 담배의 흡연자/비흡연자의 비율이 유사한 반면 궐련형과 액상형 전자담배 흡연자의 스트레스정도가 2,3점에 더욱 몰려 있는 것이 확인된다. 특히 중증환자가 눈에 띄게 줄어든 액상형 전자담배 흡연자의 스트레스가 증가한 것으로 보아 건강에 부담이 적은만큼 흡연으로 인한 만족도 역시 줄어든 것으로 사료된다. 이는 전자담배가 일반담배를 완전히 대체하기 힘든 제품임을 보여준다.

1. 회귀 모형

종속변수를 포함한 모든 변수가 독립변수이며 “스트레스정도”를 제외하면 모두 이항분포를 따르기에 로지스틱 회귀분석기법을 사용하여 모형을 추정하며 “스트레스정도”는 1을 기준으로 factor지정하였다. 파이썬의 EDA 결과를 염두에 둔 채 R 프로그램에서 Full Model을 만든 후 적정변수를 탐색하였고 P-Value가 0.05보다 낮은 값 만을 포함하여 진행하였다. 그 결과, 일반담배는 “폐암”, “신체활동\_저강도” 변수가, 궐련형 담배는 “고혈압”, “스트레스정도” 변수가, 액상형 전자담배는 “스트레스정도” 변수만이 적정변수로 선정되었다. 오즈비는 일반담배의 경우 “폐암”이 2.31배, “신체활동\_저강도”가 1.80배 증가시키고 궐련형 담배는 “고혈압” 0.67배, “스트레스정도”는 2.71배, 액상형 담배는 “스트레스정도”가 5.42배 증가시킴이 확인된다. 또 범주형 변수지만 많은 변수에 Q-Q plot 등의 그림을 그려보아 시각적으로 확인하려 하였고 역시 눈에 띄는 특징은 보이지 않았다. Adj-R^2 값이나 BIC점수 등 역시 확인하였으나 초기 확인했던 “폐기능”외에 유효한 변수는 찾지 못하였기에 Final Model 또한 초기와 같은 모델을 사용하기로 한다.

**III 결론**

• 초반 EDA에서 확인한 것과 같이 변수간의 상관관계가 크게 드러나지 않았다. 유일하게 보여지는 “폐기능”변수만이 담배와의 상관관계가 있는 것으로 보이나 본래 목적이었던 회귀모형의 구축 및 검증을 통한 담배종류간의 차이는 확인하지 못하였다. 이에 따라 사회적 통념과는 조금 다르게 담배는 종류에 상관없이 고혈압, 뇌졸증, 폐암과는 뚜렷한 인과관계가 없으며 스트레스와 폐기능에 소폭 연관성이 있다고 결정짓는다.

**IIII 한계점**

적합한 주제선정 실패가 이번 분석보고서 실패의 가장 큰 요인이라 생각한다.

1. 모든 변수를 범주형 변수로 선택하였다. 범주형 변수더라도 0과 1등의 숫자로 데이터변환을 한 후에 수치형처럼 사용할 수 있지 않을까 하는 생각으로 주제를 선정하였으나 실제로 사용하기 위해 조사해보니 현재 수준보다 더욱 복잡한 변환과정이 필요하였다.
2. 애초에 변수의 인과관계가 반대로 되어있다. 실질적으로는 흡연이 뇌졸증, 폐암 등 병을 유발시키는 원인인데 병을 원인으로 잡고 실험을 진행하였다. 흡연을 했더니 암에 걸렸다. 그러니 흡연은 암을 유발한다는 방향이 통념적이나 본 연구는 암에 걸린 사람 중 흡연자가 있으니 흡연은 암을 유발한다는 반대방향으로 진행된다. 표본이 굉장히 큰 조사였다면 역방향의 흐름도 가능했겠으나 본 연구의 데이터는 애초에 흡연자가 중심이 아니었기에 역방향 주장이 힘을 가질 수 없다.
3. 설명변수의 설정 역시 오류가 있다. 선택의 폭이 좁았으며 고혈압, 간단한 폐기능 이상 정도의 질병은 상당수의 대한민국 국민이 가진 질병이다. 흡연자가 적은 데이터에서 흡연자가 아니어도 충분히 걸리는 질병으로 연관성을 주장하기엔 설득력이 부족하다.
4. 조금 아쉬운 점은 R이 통계프로그램이 내장된 툴 이라는 이미지가 강했기에 데이터의 처리를 R로 해볼 시도 자체를 안한 것 이다. 이 후 통계를 더 깊이 공부하며 R프로그램을 사용한다면 교수님께서 수업 중 언급하셨던 dplyr이라는 패키지를 사용하여 파이썬에서 하듯이 전처리를 해보고 싶다.