# 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

• 국민건강영양조사 지침서 I.조사개요를 읽고, 본 조사의 내용을 2~3줄 분량으로 요약하여 작성합니다.

국민의 건강 및 영양 상태를 파악하기위해 실시된 본 조사는 제 8기의 조사로서 순환 표본 조사를 이용하여 1월부터 12월까지 약12개월 동안 진행 되었으며, 총 576조사수, 14,400가구 (연간 192 조사구, 4800가구)이며 조사구에서 가구의 추출 단위를 이용하여 조사를 진행하였다. 시도, 동읍면, 주택유형( 내재적층으로 성별, 연력, 주거면적, 가구주의 학력등)이 추가 되었다.

• 작성된 연구계획서(HW4)의 연구목표를 중심으로 작성하고, 앞으로 전개될 내용에 대한 소개를 작성합니다.

인슐린이 제대로 기능하지 못하여 혈당이 지나치자 높아진다. 고혈당 혈액이 온몸을 돎면서 미세혈관, 거대 혈관을 망가뜨려 당뇨병을 유발하고 각종 합병증을 유발하는 소리 없는 공포인 질병이다. 이러한 당뇨병의 근원인 인슐린과의 연결관계를 비교 분석하여 알고자 연구를 진행하게 되었다. 당뇨 환자 중 인슐린 투여군과 비 투여군의 혈당 차이를 비교하여 그 중 당뇨병 유병 여부에 따라 공복 혈당 평균 차이를 비교하여 혈당과 인슐린, 당뇨의 관계를 통계적으로 분석해 보고자 하였다.

# II 본론

• 작성된 연구계획서(HW4)의 연구내용과 연구방법의 내용에 해당됩니다.

• 먼저 연구내용과 연구방법을 소개하고, 분석하고자 하는 변수들을 소개하고, 분석 대상자들을 언급합니다.

2019년 국민건강영양조사의 당뇨병 유병여부(19세 이상), 인슐린 주사 투여, 공복 혈당, 당 화 색소를 사용할 것이며 변수는 차례대로 HE_DM_HbA1c, DE1_31, HE_glu, HE_HbA1c이다. 분석 대상자들은 당뇨병 유병 환자이면서 인슐린을 투여하는 대상으로써 결측치를 제외한 41명이다.

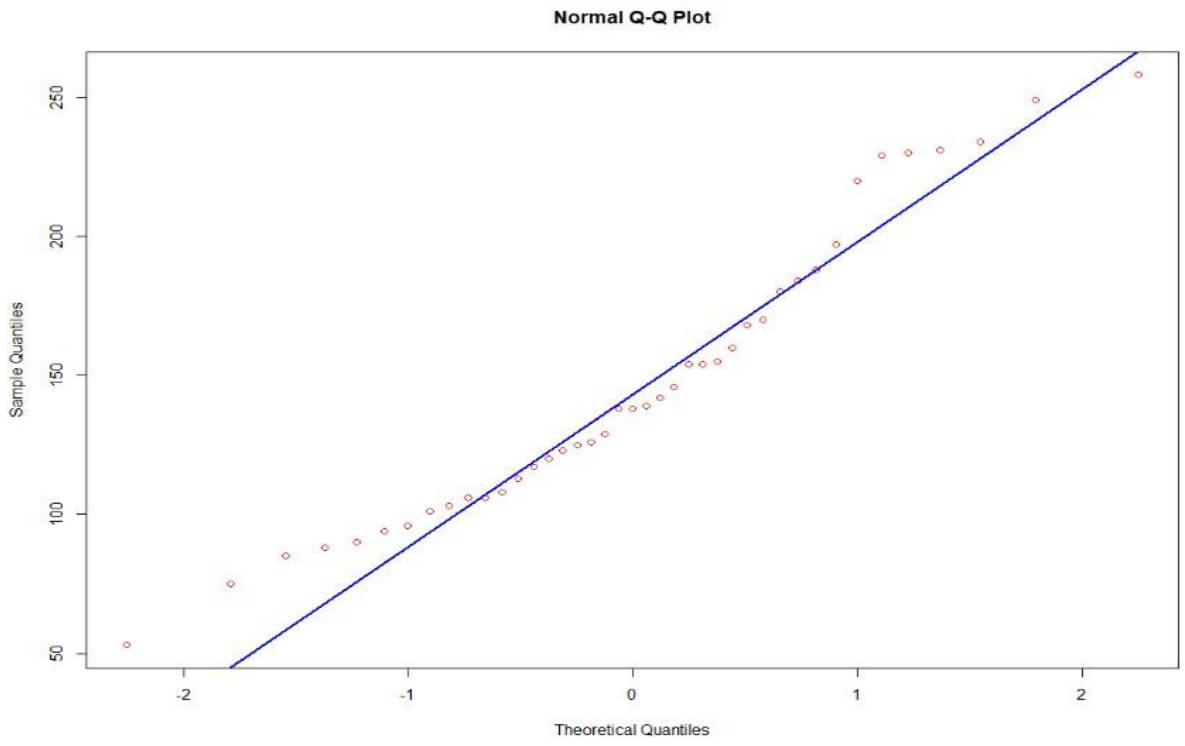
회귀 분석 전 변수를 지정하고 당뇨 유병 환자를 1로 두고 인슐린 투여 대상을 1을 subset함수를 이용하여 추출하였고 nrow 함수를 통해 최종 대상자를 확인하였다. 귀무가설은 '당뇨병 유병여부에 따라 공복 혈당 평균의 차이가 없다' 라고 가정하였다. 그후 정규성을 확인하며 회귀분석을 진행하였다.

• 분석은 탐색적 분석과 회귀모형으로 나누어 분석을 진행합니다.

① 탐색적 분석

- 회귀분석 전, 변수들간의 상관관계를 산출하고, 필요시 산점도를 삽입합니다.

(단, 변수가 많은 경우, 모든 산점도를 넣지 말고, 뚜렷한 선형관계가 있는 경우만 고려합니다.)




공복혈당 (HE_glu)에 대한 Q-Q plot을 이용한 산점도이다. 도트로 표현된 분포는 누적 분포함수, 즉 우리가 가진 공복 혈당의 분포이고, 파란색 선으로 표현된 분포는 정규분포를 나타낸다.

이 두 분포가 근접하게 분포하면 정규성을 띈다고 하고 근접하지 않으면 정규성을 띄지 않는다고 정의한다. 위의 산점도는 두 분포가 근접하게 분포하지 않아 정규성을 띄지 않는다.

② 회귀 모형

- 탐색적 분석의 내용을 바탕으로 적절한 회귀모형과 모형에 대한 검정의 내용을 작성합니다.

정규성 검정을 위해 Shapiro-wilk 검정을 이용하였다

> H0: 정규분포를 따른다. vs. H1: 정규분포를 따르지 않는다.

```
        Shapiro-Wilk normality test

data:  data_subj$HE_glu
W = 0.95058, p-value = 0.07336
```

정규성 검정 결과 p-value<0.05이므로, 귀무가설을 기각하여 정규분포를 따르지 않는다.

다음은 등분산성 검증이다.

> H0: 두 그룹의 분산의 차이가 없다. vs. H1: 두 그룹의 분산의 차이가 있다.

```
> var.test(x,y)

        F test to compare two variances

data:  x and y
F = 0.22004, num df = 149, denom df = 149, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1594015 0.3037352
sample estimates:
ratio of variances
         0.2200361
```

두 번째 줄에 등장하는 p-value는 매우 작으므로 귀무가설을 기각하여 등분산성이 없다.

즉, 두 그룹의 분산이 같지 않다고 할 수 있다.

- 정규 분포를 따르지 않고, 인슐린 투여 그룹 간의 분산이 다르다.

t-test

t- test는 정규성을 만족해야하며 등분산성을 알아야한다. 앞의 결과에서 정규성을 만족하지 않는다고 하였지만 우리는 t-test진행을 위해 정규분포를 따른다고 가정하고 분석을 진행하였다.

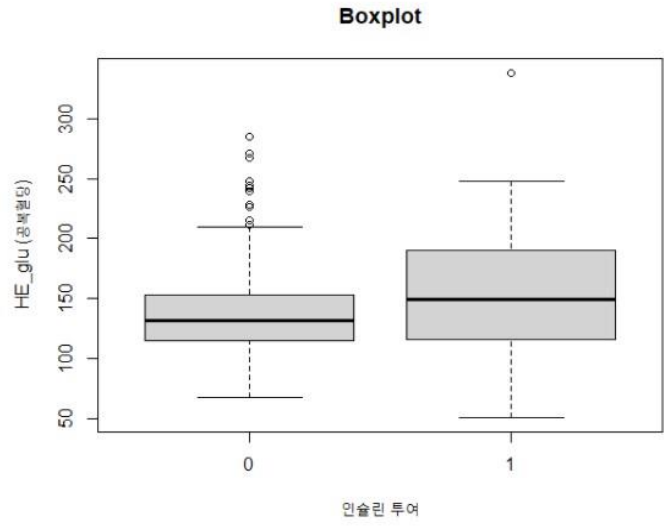
```
> t.test(x,y)

        Welch Two Sample t-test

data:  x and y
t = 13.098, df = 211.54, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.771500 2.399166
sample estimates:
mean of x mean of y
 5.843333  3.758000
```

등분산성 여부를 만족하지 못했기에 var.equal=F를 추가하고, 신뢰구간은 일반적으로 95%로 설정하여 level=0.95까지 추가하여 진행하였다. 실행 결과 p-value는 2.2e-16보다 작으므로 인슐린 투여 여부에 따라 공복 혈당이 다르지 않다고 할 수 있다.

# III 결론

- 본론에서 얻은 결과를 바탕으로 내용을 정리합니다.

공복 혈당은 평균을 구할 변수이고 인슐린 투여는 그룹 변수로 설정하여 진행하였고 정규성 검정 결과 인슐린 투여 여부에 따라 공복혈당이 다르지 않다는 것을 알 수 있다. 위의 t-test검정 결과에서 x인 인슐린 비투여의 평균은 5.843333, y인 인슐린 부여군의 평균은 3.758000임을 확인 할 수 있다. 평균의 차이가 있어 보이는데 인슐린 투여여부와 공복혈당이 다르지 않게 결과가 도출되었다. Box-plot 이용하여 인슐린 투여여부에 따른 공복 혈당을 비교해 보았다.



두 표본 평균의 차이가 나는 것을 확인할 수 있지만, 평균의 차이에 비해 분산이 커서 명확한 차이가 드러나지 않는다. t-test에 사용되는 t 통계량은 분산까지 고려한 평균의 차이이기 때문이다. 그렇기에 분산이 큰 경우에는 평균의 차이가 드러나지 않게 되었다.

@ 결론 적으로 인슐린 투여 여부에 따라 공복 혈당이 크게 다르지 않아 당뇨 환자에게 인슐린 투여군과 비 투여군의 혈당의 차이는 크게 나지 않았다.

- 아쉬웠던 점이나 추후 더 살펴보고자 하는 내용이 있다면 함께 기입합니다.

분산 값에 따른 평균의 차이에 따라 전체적인 가설의 명확한 결론이 나지 않은 것 같아 아쉬운 점이 있다. 또 예상된 결론과 조금 달라 다른 가설을 제시해 더욱 명확한 당뇨병과 다른 합병증과의 관계를 추후 더 살펴보고 싶다. 고혈압, 당뇨를 제외한 중증 질환의 유전관계도 살펴볼 예정이다. 회귀분석 자료조사를 배우며 경증 당뇨를 앓고 계시는 아버지에게 도움이 될 수 있을 거 같아 뿌듯하기도 하다.