**HW1 (Due – 10/11)**

**222dbg04 이지은 빅데이터분석학**

**Note:** You can ask friends for help using SAS, reviewing lecture, and discussing strategies to address the questions. But the analysis plan, results description, and development of your table should be entirely your own.

**Answer the following questions using HW1.xlsx file [Scope:** materials covered in class]

**--------------------------------------------------------------------------------------------------------------------------------------**

**Data dictionary**

**Gender\_n:**

0 = Female

1 = Male

**Age**: age in years

**Height**: height (meters)

**Weight**: weight (kg)

**Family\_history\_with\_overweight\_n**:

1= Yes

0= No

**FAVC\_n**: Frequent consumption of high caloric food (Do you eat high caloric food frequently?)

1= Yes

0= No

**FCVC**: Frequency of consumption of vegetables (Do you usually eat vegetables in your meals?)

(continuous)

**NCP**: Number of main meals (How many main meals do you have daily?)

(continuous)

**CAEC\_n**: Consumption of food between meals (Do you eat any food between meals?),

0= Always

1= Frequently

2= Sometimes

3= no

**SMOKE\_n**: do you smoke?

1= Yes

0= No

**CH2O**: Consumption of water daily (How much water do you drink daily?)

(continuous)

**CALC\_n**: Consumption of alcohol (how often do you drink alcohol?)

0= Always

1= Frequently

2= Sometimes

3= no

**SCC\_n**: Calories consumption monitoring (Do you monitor the calories you eat daily?)

1= Yes

0= No

**FAF**: Physical activity frequency (How often do you have physical activity?)

(continuous)

**TUE**: Time using technology devices (how much time do you use technological devices such as cell phone, videogames, television, computer and others?)

(continuous)

**MTRANS\_n**: Transportation used (Which transportation do you usually use?)

0= Automobile

1= Motorbike

2= Bike

3= Public Transportation

4= Walking

You are interested in answering the following question “ is age associated with bmi ?”

Be sure to justify the reasons and necessity for using the statistical procedures that you use; Be sure to provide effect estimates and enough detail so that I can evaluate your work.

1. Study population
   1. Define exclusion/inclusion criteria to address the research question and justify the reason why you set those criteria.

**성인남녀 키와 몸무게가 0이상이며, 그 외의 변수들이 결측 값이 아닌 집단. 결측 값의 영향을 제거하고 타요인의 영향을 받을 수 있는 청소년/유아기의 집단은 제외하였습니다.**

* 1. Apply your criteria reading the dataset into sas. After application of your inclusion/exclusion criteria, what is your total study sample size?

**1438명**

**/\*코드\*/**

**proc import**

**datafile="C:\Users\USER\Desktop\HW1.xlsx"**

**out=saslab**

**dbms=xlsx**

**replace;**

**GETNAMES=YES;**

**run;**

**data filtered\_data;**

**set Saslab;**

**/\* 포함 기준 \*/**

**if Gender\_n in (0, 1) and Age >= 20 and Age <= 100 and Height > 0 and Weight\_2 > 0**

**and FCVC >= 0 and SMOKE\_n in (0, 1) and CH2O >= 0**

**and TUE >= 0 and Family\_history\_with\_overweight\_n in (0,1) then**

**output;**

**run;**

**proc means nmiss min p50 max;/\*결측값 확인\*/**

**var Gender\_n Age Height Weight\_2 Family\_history\_with\_overweight\_n FAVC\_n FCVC NCP CAEC\_n SMOKE\_n CH2O SCC\_n FAF TUE MTRANS\_n;**

**run;**

**/\* 전체 샘플 수 계산 \*/**

**proc sql;**

**select count(\*) as sample\_n**

**from filtered\_data;**

**quit;/\*샘플수: 1438\*/**

1. Confounding theory
   1. What is a confounding factor ?

**confounding factor란 교란변수, 즉 보정이 필요한 변수를 의미한다.**

**교란은 결과와 연관된 요소들의 불균형으로 두 개 이상의 요인 효과가 혼합되어 분리할 수 없는 것을 의미한다.**

* 1. Define the three conditions of confounding factors

**교란의 3가지 조건은 다음과 같다. 1.Exposure과 연관이 있다. 2. Outcome과 연관이 있어야 한다. 3. 변수가 intermediate하게(중간요인으로) pathway중에 위치하면 안된다.(Exposure에 밀접한 영향을 주면 안된다)**

* 1. Why do we have to adjust for confounding factor ?

**보정을 하지 않으면 true association을 포착할 수 없기 때문에 교란 요인을 보정해야 한다.**

1. Selection of confounding factor
   1. Describe three strategies of selection of confounding factors and discuss pros and cons

**1)선행탐색: risk factor를 키워드로 기존 연구 리뷰를 통해 일반적으로 잘 알려진 교란요인을 탐색한다. 가장 우선되는 방법이지만 희귀 질환 같이 사례가 없는 경우에는 선행연구를 통한 탐색이 불가능하다. 이 경우 데이터 탐색을 통해 진행한다.**

**2)데이터 탐색: 데이터를 탐색하여 선택된 risk factor의 exposure category에 따른 분포의 차이를 확인한다. 그러나 특정 데이터셋에서만 우연히 포착되는 상관관계일 수 있으므로 오버피팅의 가능성이 존재한다. 따라서 실제 인과성이 있는 risk factor인지 biological mechanism적으로 다시 생각해볼 필요가 있다.**

**3)10% estimate change rule에 기반하여 10% 이상의 변화가 없을 경우, 해당 교란 변수를 모델에서 제외한다. 즉, 잠재적인 교란 요소를 보정하는 것이 crude association을 10%이상 변화 시킨다면, 교란변수로 지정한다.**

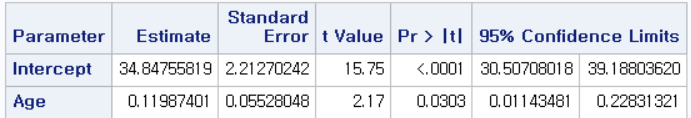
* 1. Try to apply those strategies to address your research question “ is age associated with bmi ?” in your study population you set in question 1
     1. Methods 1

# Several studies have found an association between obesity risk and urban sprawl, a pattern of development characterized by decentralization, low density land uses, and **automobile-focused transportation systems** ([9](https://onlinelibrary.wiley.com/doi/full/10.1038/oby.2007.251#b9), [10](https://onlinelibrary.wiley.com/doi/full/10.1038/oby.2007.251#b10), [11](https://onlinelibrary.wiley.com/doi/full/10.1038/oby.2007.251#b11), [12](https://onlinelibrary.wiley.com/doi/full/10.1038/oby.2007.251#b12)).

Research suggests **that the accessibility of affordable, high quality food** is associated with better diets and nutrition ([20](https://onlinelibrary.wiley.com/doi/full/10.1038/oby.2007.251#b20)). These types of foods are more likely to be found in supermarkets than in convenience stores, and people who eat meals prepared at home, presumably made from foods purchased in supermarkets rather than purchased pre-made at a takeout restaurant or consumed at a restaurant, were more likely to have a reduced obesity risk ([21](https://onlinelibrary.wiley.com/doi/full/10.1038/oby.2007.251#b21)).

# Individual variables, previously found to be or hypothesized to be associated with obesity risk, were provided by the Mass DPH. **These included age, education, income, race/ethnicity, smoking status, and sex**. [출처: Neighborhood Risk Factors for Obesity]

**Method1의 교란 변수: Gender, SMOKE\_n, FCVC, FAVC\_n, MTRANS\_n**



이때, age변수의 p-value가 0.05보다 작은 0.03이므로 age는 bmi에 유의한 변수라고 할 수 있다.

* + 1. Methods 2

**/\*Outcome\*/**

**proc glm; model bmi = age/solution clparm; run;**

**proc glm; model bmi = Height/solution clparm; run;**

**proc glm; model bmi = Weight\_2/solution clparm; run;**

**proc glm; model bmi =FCVC/solution clparm; run;**

**proc glm; model bmi = NCP/solution clparm; run;**

**proc glm; model bmi = CH2O/solution clparm; run;**

**proc glm; model bmi = FAF/solution clparm; run;**

**proc glm; model bmi = TUE/solution clparm; run;**

**proc glm; class Family\_history\_with\_overweight\_n; model bmi = Family\_history\_with\_overweight\_n/solution clparm; run;**

**proc glm; class FAVC\_n; model bmi = FAVC\_n/solution clparm; run;**

**proc glm; class CAEC\_n; model bmi =CAEC\_n/solution clparm; run;**

**proc glm; class SMOKE\_n; model bmi =SMOKE\_n/solution clparm; run;**

**proc glm; class SCC\_n; model bmi = SCC\_n/solution clparm; run;**

**proc glm; class MTRANS\_n; model bmi = MTRANS\_n/solution clparm; run;**

**/\*Exposure\*/**

**proc glm; model age= Height/solution clparm; run;**

**proc glm; model age = Weight\_2/solution clparm; run;**

**proc glm; model age =FCVC/solution clparm; run;**

**proc glm; model age = NCP/solution clparm; run;**

**proc glm; model age = CH2O/solution clparm; run;**

**proc glm; model age = FAF/solution clparm; run;**

**proc glm; model age = TUE/solution clparm; run;**

**proc glm; class Family\_history\_with\_overweight\_n; model age = Family\_history\_with\_overweight\_n/solution clparm; run;**

**proc glm; class FAVC\_n; model age = FAVC\_n/solution clparm; run;**

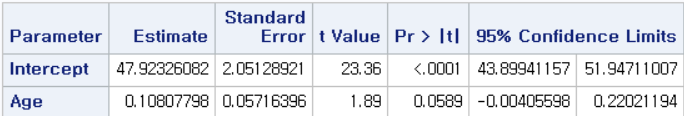
**proc glm; class CAEC\_n; model age =CAEC\_n/solution clparm; run;**

**proc glm; class SMOKE\_n; model age =SMOKE\_n/solution clparm; run;**

**proc glm; class SCC\_n; model age = SCC\_n/solution clparm; run;**

**proc glm; class MTRANS\_n; model age = MTRANS\_n/solution clparm; run;**

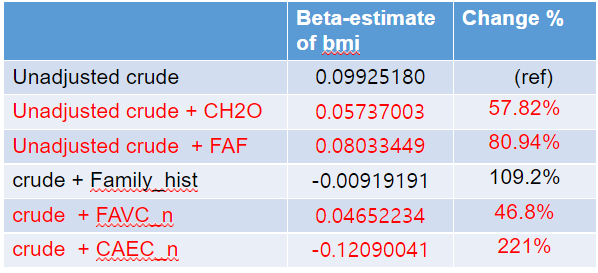
**Method2의 교란 변수= CH2O, FAF, MTRANS\_n, CAEC\_n, FAUC\_n, Family\_hist**

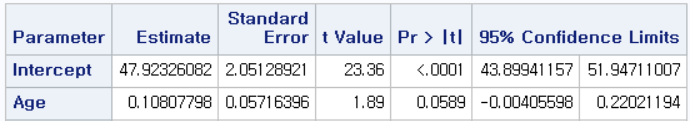


* + 1. Methods 3

**10% rule에 기반하여 교란변수 지정.**

**Method 2-2의 교란 변수= CH2O, FAF, CAEC\_n, FAVC\_n**





1. Among three methods of selecting confounding factors, what methods would you choose ? why ? justify your answer briefly.

사실 일반적으로 주제를 잡고 교란 요인을 결정할 때는, 주제에 대한 선행 연구를 통해 일반적인 교란 요인을 확인하고 일반적인 교란 요인이 정해져 있지 않을때는 Method2를 거쳐 change 10% rule을 기반으로 교란 요인을 결정지을 것 같다. 하지만 과제의 주제인 obesity에 대해서는 일반적인 교란 요인이 나와있기 때문에(Method1에서 언급) 선행 연구를 통해 교란 요인을 정할 것 같다.

참고 문헌)

[1]R. P. Lopez, “Neighborhood Risk Factors for Obesity\*,” *Obesity*, vol. 15, no. 8. Wiley, pp. 2111–2119, Aug-2007.