**Exam 2**

**222DBG04 이지은 빅데이터분석학**

**Theoretical questions**

Q1. Let’s assume that you are interested in examining the association between saturated fat intake and BMI (continuous). In your dataset, there are 4 different variables relevant to saturate fat intake.

**Sat: continuous raw data**

**Sat\_q = quartile ordinal categorical variable for sat\_q**

**Sat\_md\_q = median saturated fat intake values assigned within each group defined by quartiles of sat\_q**

**BMI = bmk, kg/m2 (continuous)**

1. Write down the equation model [*e.g., E[BMI| X1, X2,X3,,,] = a0 + a1\*X1 + a2\*X2 +a3\*X3…*] you need to develop to present results of comparing mean of BMI across groups defined by quartiles of saturated fat intake.

**=> *E[BMI|*  Sat\_q*]= a0 + a1\** Sat\_q *+ a2\** Sat\_q *+a3\** Sat\_q**

1. What would be the rationale using the model you are specifying in Q1-a) over developing a model including a one term of continuous raw variable of saturated fat intake (sat) as an independent variable ?

**=> 포화 지방 섭취가 어떻게 BMI에 영향을 미치는지 파악하기 위해서 Sat변수를 독립변수로 방정식에 추가하였다.**

1. Explain what the p-trend test is.

**=> 범주형 독립 변수와 연속형 종속 변수 간의 경향을 테스트하는 통계 검정 방법. 이 검정은 범주형 변수의 각 범주가 종속 변수에 대해 영향을 주는지 검정한다..**

1. Write down the equation model conducting a p-trend test using “sat”, “sat\_q”, and “sat\_md\_q”. Then, discuss pros and cons of each trend test.

**=> E[BMI|X1, X2, X3] = a0 + a1\*Sat + a2\*Sat + a3\*Sat**

**=> E[BMI|X1, X2, X3] = a0 + a1\*Sat\_q + a2\*Sat\_q + a3\*Sat\_q**

**=> E[BMI|X1, X2, X3] = a0 + a1\* Sat\_md\_q + a2\* Sat\_md\_q + a3\*Sat\_md\_q**

**Sat: 포화 지방 섭취의 연속적인 변화에 따른 BMI의 선형적인 추세를 확인할 수 있으며, 정보 손실 없이 모든 데이터를 사용할 수 있다. 또한 가장 강력하고 효율적인 방식이다. 하지만 아웃라이어에 취약하며 선형성이 만족되어야 한다는 단점이 있다.**

**Sat\_q: 포화 지방 섭취를 여러개의 범주로 나누어 각 범주별 평균 BMI를 비교함으로써 비선형 효과를 파악하고자 하는 방식이다. 위보다 아웃라이어에 덜 민감한 방식이다. 하지만 올바른 컷오프 값을 선정해야 하고 더 적은 정보의 데이터를 사용하므로 효과가 비교적 약하다는 단점이 있다.**

**Sat\_md\_q: 사분위수 범주 내에서 중앙값을 사용하여 범주를 생성하는 방식이다. 데이터의 분포를 반영하고 이상치에 덜 민감하다. 하지만 데이터 손실이 존재할 수 있고 연속형보다 효과가 약하다는 단점이 있다.**

Q2. Read article by Cantrol et al (Anticancer Research, 2019; <https://ar.iiarjournals.org/content/anticanres/39/11/6197.full.pdf>).

1. Read study population section and answer the following
   1. What were the inclusion/exclusion criteria for the study population selection ? What do you think were the rationale for the inclusion/exclusion criteria ?

**no contraceptive or replacement hormones taken in the 3 months preceding mammography, no Tamoxifen® or Raloxifene® treatment, no current pregnancy or breastfeeding, no personal history of cancer (all sites combined), no breast reduction or augmentation, and no endocrine diseases. There was no age limit, and menopausal status was defined according to the criteria of the Nurses’ Health Study (16) based on each woman’s age, menstrual cycles, smoking status and history of gynaecological surgery. Women with intellectual disabilities were ineligible for the study because of their inability to provide informed consent. Of the 1574 women (783 premenopausal and 791 postmenopausal) who participated in the original study, we excluded eleven women who did not complete the food frequency questionnaire (FFQ), three women who reported extreme daily caloric intake (5000 kcal/day), four women for whom the waist-to-hip ratio was not available, two for missing data on physical activity and eight who did not fill out their hormone use. 이러한 기준으로 최종적으로 1546명이 연구에 참여하게 되었다. 이러한 inclusion/exclusion criteria가 연구에 간접적인 영향을 주는 요인을 모두 배제하였으므로 적합한 기준이라고 생각한다.**

1. Read Statistical methods section of Cantrol et al (Anticancer Research, 2019). Briefly summarize the statistical methods relevant for table 2
   1. How was dietary exposure modelled in the study? What are pros and cons of modelling the exposure variables as did in the study ?

**=> 유제품의 소비 빈도는 각 카테고리의 중간을 사용하여 주당 섭취량으로 변환하였다. DF의 종류(우유, 아이스크림, 요구르트, 치즈)와 지방성분(저지방, 고지방)에 따라 15개 유제품군을 구성하고 총 섭취빈도, 저지방, 고지방 DF섭취빈도 등을 합산한 변수를 작성하였다.**

**이러한 모델링은 섭취 빈도를 주당 섭취량으로 변환함으로써 유제품 섭취의 정확한 추정을 제공한다.**

**유제품 유형과 지방 종류에 따른 변수 변환은 유제품이 유방 밀도에 미치는 영향을 더 반영하여 분석할 수 있다. 하지만 분위수로 범주화하는 것은 정보의 손실을 유발할 수 있다는 단점이 있다.**

* 1. How were the confounding variables selected in the study? What is the pros and cons of the confounding selection methods used in the study ?

**=> Potential confounders가 선험적으로 선택되어 회귀와 상관관계 모델에 적용되었다. 다음과 같은 변수들을 선택하였다. age (years), body mass index (kg/m2), waist-to-hip ratio, menarche age (years), age at first full-term pregnancy (years), number of term pregnancies, duration of breastfeeding (years), duration of oral contraceptive use (years), duration of hormone replacement therapy use (years), family history of breast cancer**

**(yes/no), personal history of breast biopsy (number of biopsies),**

**smoking (non-smoker, former or current smoker), education**

**(elementary or less, high school, college or university), physical**

**activity (MET-h/week), alcohol consumption (drinks/week), total**

**caloric intake (kcal/day), and menopausal status when applicable**

**이렇게 사전에 confounding factors를 정의하면 선행 연구를 통해 널리 알려진 요인들을 고려할 수 있다. 하지만 bias나 다중공선성의 문제가 발생할 수 있으며, confounding 요인에 대해 정확한 선택이 어려울 수 있다.**

* 1. How was the trend test conducted in the study ? What are the pros and cons of the trend test methods used in the study ?

**=>연구에서는 유제품 소비의 분위에 따른 BD 평균을 추정하고 사분위수에 trend testrk 수행되었다. 이 방법을 통해 선형 추이를 통해 잠재적인 exposure과 output사이의 관계를 살필 수 있다. 하지만 이런 trend test는 선형성을 가정하기 때문에 선형성을 만족하지 못한다면 설명력이 감소할 수 있다.**

* 1. Write down 4 model equations used to produce the results of total milk consumption and breast density in Table II
     1. Unadjusted model : E[BD | total milk consum] = Breast Density = b0 + b1 \* Total Milk Consumption
     2. Unadjusted trend model: Breast Density = b0 + b1 \* Total Milk Consumption
     3. Adjusted model:
     4. Adjusted trend model

1. Interpret the results of high-fat ice cream consumption in relation to percent breast density.

**=>표2를 참고하면 고지방 아이스크림 섭취의 p trend값이 0.043으로 0.05미만이다. 따라서 해당 변수는 통계적으로 유방 밀도에 유의한 영향을 준다고 결론 내릴 수 있다.**

Q3. What is the impact of missing data to the study results ? In other words, why is addressing missing data important while conducting biomedical analyses ?

**=> 데이터의 폐기로 인한 정보의 손실은 샘플 사이즈를 감소시키고 완전한 측정값의 결과만 도출되게 된다. 또한 bias가 결과에 발생하게 될 수 있다.**

**분석에서는 완전한 관측치만 사용하기 때문에 샘플사이즈가 감소하고 그 특정 관측치들이 전체 모집단을 반영한다고 확신할 수 없기 때문에 문제가 발생한다. 이러한 문제점 때문에 결측치를 처리하게 된다.**

Q4. Discuss the four methods of missing handling technique except IPW methods. What are pros and cons?

**-complete case method: 결측값이 있는 모든 값을 삭제하는 방식. 결측치의 수가 적을수록 효율적이다. 간단하고 쉬운 방법이지만 MCAR가 아닌 경우 bias된 결과 도출하게 된다는 단점이 있다.**

**-simple imputation: 결측값을 평균이나 예측값, 랜덤값 등으로 대체하는 방법. 데이터의 수를 감소시키지 않아 설명력이 감소하지 않지만 결측치를 특정값으로 대체하는 것이 괜찮은지에 대한 불확실함이 존재한다.**

**-Missing indicator methods: 결측이면 1, 결측이 아니면 0을 가지는 인디케이터 변수를 생성한다. 데이터를 버리는 것보다 더 효과적이지만 여전히 불확실성이 존재하며, 상관관계를 확인하기 어렵다는 단점이 있다..**

**-Multiple imputation: 각 개개인의 결측마다 2개 이상의 다양한 값으로 대체하는 방식이다.**

**예를 들어 10명이 결측이면, 각각의 데이터셋에서의 모델 결과값으로 풀링하여 추정한다. 연산도 많이 필요하고 복잡한 방법이기 때문에, 특정 환경에서는 어려울 수도 있다. 파라미터를 반영하는 과정에서 수집가의 배경지식이 포함되기도 한다.**

**Data exercise**

1. You are interested in examining the association between vitamin A intake and BMI (bmi1, continuous term). Delete population without BMI or vitamin A data. Read in cohort\_all\_eng.sas7bdat. Then, examine the association between vitamin A intake and BMI.
2. Complete the table below

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Tertile of vitamin A | | |  |  |
|  | T1 | T2 | T3 | P-trend |  |
| In all |  |  |  |  |  |
| Crude model | (ref) | 0.19170990 | 0.35806801 | **0.0014** |  |
| MV model | (ref) | 0.16791459 | 0.36292917 | **<.0001** |  |
|  |  |  |  |  |  |

a.1 What is your study population selection criteria ? what was your rationale ?

=>성인이며 심혈관질환과 당뇨병을 앓고 있지 않은 population을 선정하였다.

a.2 What is your strategy for selecting confounding factors ?

=>선행 연구에서 confounding factor들을 탐색하였다.

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]

따라서, edu\_grp, income\_grp, smoke\_cu, sex, age들을 리스크팩터로 두고 분석하였다.

a.3 What is your strategy for addressing missing data ?

=> 결측값을 모두 제거하여 사용하였습니다

a.4. What is your strategy for conducting trend test ?

=>vita변수를 구간을 3개의 그룹으로 나누어 **각 범주별 평균 vita를 비교함으로써** trend test를 진행하였다.

**proc** **rank** out= trend groups=**3**;

var vita;

ranks vita\_q; **run**;

**proc** **glm**;

class vita\_q(ref='0');

model BMI1=vita\_q/solution clparm;**run**;

a.5. Write down your model equation and SAS code for crude tertile vitamin A model and crude vitamin A p-trend model, MV tertile vitamin A model and MV p-trend model

=>

BMI = b0 + b1\*vita

BMI = b0 + b1\*vita\_q

BMI = b0 + b1\*vita + b2\* edu\_grp + b3\* income +b4\* smoke\_cu+b5\* sex +b6\*age

BMI = b0 + b1\*vita\_q + b2\* edu\_grp + b3\* income +b4\* smoke\_cu+b5\* sex +b6\*age

/\*sas code\*/

**proc** **glm**;

class vita\_q(ref='0');

model BMI1 = vita\_q / solution clparm;

**run**;

**proc** **glm**;

model BMI1 = vita / solution clparm;

**run**;

**proc** **glm**;

class vita\_q(ref='0');

model BMI1 = vita\_q edu\_grp income\_grp smok\_cu sex age / solution clparm;

**run**;

**proc** **glm**;

model BMI1 = vita edu\_grp income\_grp smok\_cu sex age / solution clparm;

**run**;

1. What do you conclude between vitamin A intake and BMI in all population based on the results above ? Why or Why not ?
2. => p trend값이 0.05 미만이므로 비타민 A 섭취량이 BMI에 통계적으로 유의한 영향이 있는 것으로 결론 내릴 수 있다.

c. Examine the association between vitamin A intake and BMI in all population and population defined by alcohol drink status (drink\_cu).

c.1 Complete the table below

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Tertile of vitamin A | | |  |  |
| By alcohol status\* | T1 | T2 | T3 | P-trend | P-interaction |
| No current | (ref) | -0.29059925 | -0.06756600 | 0.0082 | 0.3102 |
| Current | (ref) |  |  |  |  |
|  |  |  |  |  |  |

c.2 Write down sas code you used to complete the table C.

**proc** **glm** ;

class drink\_cu vita\_q(ref='0');

model BMI1 = drink\_cu vita\_q drink\_cu\*vita\_q/solution clparm;

**run**;

c.3 how did you calculate the p-interaction

d. Given the results above, is the association between vitamin A and BMI different between current alcohol drinker vs. no drinker ? Why or Why not ?

=> p trend값이 0.05 미만이므로 음주여부가 비타민a의 섭취가 BMI에 미치는 상관성에 영향이 주는 것으로 결론 내릴 수 있다.

Data dictionary

|  |  |  |
| --- | --- | --- |
| Variable name | Data type | Variable Dictionary |
| id | 문자 | Participants ID |
| edu\_grp | 숫자 | Education  1= <high school; 2 = Undergraduate; 3= ≥ graduate school |
| income\_grp | 숫자 | Income level  0 = very low,1=low,2=medium,3=high |
| bmi1\_grp | 숫자 | Overweight status  0 = underweight,1=normal weight,2= overweight,3= obese |
| drink\_cu | 숫자 | Current drinking status  0= no,1= current |
| smok\_cu | 숫자 | Current smoking status  0=no, 1=current |
| sex | 숫자 | Gender  0= male,1= female |
| energy | 숫자 | Energy intake/d (Continuous), kcal |
| age | 숫자 | Age (Continuous), years |
| vita | 숫자 | vitamin A intake (Continuous), mcg/d |
| DM | 숫자 | Diabetes status  0 = no,1= yes |
| BMI1 | 숫자 | BMI, kg/m2 |
| na | 숫자 | Sodium intake, continuous |
| CVD | 숫자 | CVD disease,  0= no, 1= yes |
| Vitc\_n | 숫자 | Vitc intake, continuous |
| Coffee\_fn | 숫자 | Coffee intake, continuous |
| VITA\_4Q | 숫자 | Vit A intake, quartile |
| NA\_4Q | 숫자 | Sodium intake, quartile |
| vitc\_n\_4Q | 숫자 | Vitc intake, quartile |
| coffee\_fn\_4Q | 숫자 | Coffee intake, quartile |