**Data Science in Healthcare**

**Problem Set #2**

이름: 김연주

학과: 통계학과

학번: 2021250461

**PART A: Conceptual Questions**

**1. We are examining the relationship between health care spending and health outcomes. Explain all possible relationships and illustrate them in a figure, with the X-axis representing health care spending and the Y-axis representing health outcomes.**

(1) overall patient outcomes: 도표, 라인, 스케치, 그래프이(가) 표시된 사진

자동 생성된 설명출처: 강의안

의료비 지출이 0일 때는 환자의 상태가 0에 가깝고 의료비 지출이 조금 증가하면 환자의 상태가 빠른 속도로 완화된다. 그러나 일정 수준 이상부터는 의료비 지출에 따른 환자의 상태가 유의미한 변화를 갖지 않는다. 즉, 그래프상 B지점까지의 의료비 지출이 의미 있는 지출이라고 볼 수 있다.

(2) 기대수명: 출처: <https://ourworldindata.org/grapher/life-expectancy-vs-health-expenditure> 텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명

위의 1번 그래프와 마찬가지로 의료비 지출에 따른 기대수명은 초반에는 급격하게 증가하지만, 일정 수준 이상부터는 큰 변화를 보이지 않는다.

(3) child mortality: 출처: <https://ourworldindata.org/grapher/per-capita-total-expenditure-on-health-vs-child-mortality> 텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명

의료비 지출이 증가하면 아동 사망률은 감소하는 추세를 보인다.

(4) QALY (건강수명): 출처: <https://ourworldindata.org/grapher/healthy-life-expectancy-vs-health-expenditure-per-capita>

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

건강수명은 의료비 지출이 증가함에 따라 증가하는 추세를 보이지만, 변화가 급격하지 않고 완만함을 알 수 있다.

(5) DALY (질병부담): 출처: <https://ourworldindata.org/grapher/disease-burden-vs-health-expenditure-per-capita>

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

의료비 지출이 증가함에 따라 질병 부담은 줄어든다.

**2. We are evaluating health outcomes. Please choose four different measures, explain what each one is, and describe how they differ from one another.**

1. QALY (Quality adjusted life year): QALY는 건강 상태의 질을 고려하여 추가된 생명을 평가하는 지표이다. 특정 건강 상태에서 살게 될 추가 수명에 삶의 질 가중치 (preference score)를 곱하여 산출한다. 예를 들어 0.8의 선호도 점수를 지니고 5년을 더 산다면 해당 기간은 4QALY로 계산된다. 또한, 10년을 더 산다고 해도 0.3의 선호도 점수를 지니고 있다면 이는 3QALY로 앞선 경우보다 점수가 낮다. 이는 수명 연장과 삶의 질을 모두 고려하여 특정 의료 개입의 총체적 효과를 평가할 수 있기에 유용한 지표이다.
2. DALY (disability adjusted life year): DALY는 조기 사망과 장애로 인해 손실된 건강한 수명을 합산한 것으로, 질병 부담을 정량화하는 지표이다. 이는 조기 사망으로 인한 손실 연수 (YLL)과 장애로 인해 건강하지 않은 상태로 지낸 연수 (YLD)를 합한 값이다. 이는 특정 질명이나 장애가 인구 집단에 미치는 부담을 비교하고, 공중보건 자원 배분의 우선순위를 결정하는 데 유용한 지표로 쓰인다. DALY는 QALY와는 반대로 질병과 장애로 인한 부정적인 영향을 측정하는 지표이다.
3. YLL (Years of life lost due to death): 조기 사망으로 인해 손실된 수명을 측정하는 지표이다. 이는 기대 수명에 비해 일찍 사망한 연수를 계산한다. 즉, x연도에서 사망한 수에 기대수명에서 사망 연령을 뺀 값을 곱한 값이다.
4. YLD (Years of life lost due to disability): 장애나 만성 질환으로 인해 건강하지 않은 상태에서 살아가는 기간을 측정한다. 이는 질병의 중증도와 장애가 지속되는 기간을 반영한다. 즉, 특정 질명으로 인한 삶의 질 저하 정도와 장애가 지속된 기간을 곱하여 계산된다.

**3. We are calculating the YLLs, YLDs, and DALYs in the following scenarios. Consider the optimal life expectancy to be 88 years. Use the disability weights table provided below for the calculation of YLDs.**

**a. Fifty individuals experience mild depression at the age of 36, lasting two years, but live until the age of 95.**

1. YLL: 기대 수명보다 사망 연령이 높기에 YLL은 0이 된다.
2. YLD: mild depression의 장애 가중치는 0.159이다. 지속 기간이 2년이기 때문에 YLD는 0.159\*2\*50=15.9이다.
3. DALYs: YLL과 YLD의 합인 15.9가 된다.

**b. Seventy individuals contract a severe infectious disease at the age of 2 lasting 1 year. At the age of 22, they develop a cocaine dependence lasting for 5 years. Following the dependence, they are affected by a moderate depression for 8 years, which progress into a severe depression lasting 2 years. At the age of 45, they die of a heart failure.**

1. YLL: (88-45)\*70=3010
2. YLD: severe infectious disease의 장애 가중치는 0.21, cocaine dependence의 장애 가중치는 0.376, moderate depression의 장애 가중치는 0.406, severe depression의 장애 가중치는 0.655이다. 그에 따른 YLD는 70\*0.21\*1+70\*0.376\*5+70\*0.406\*8+70\*0.655\*2=464.56이다.
3. DALYs: YLL과 YLD를 합한 3474.56이 된다.

**4. We are calculating life expectancy using a life table. The Excel file contains death rate information for each age group, categorized by sex (male and female). Please follow the outlined process to complete the life table in the Excel file and estimate life expectancy overall (both) and by sex (male and female).**

엑셀 파일 첨부. Overall, male, female에 대한 기대수명이 각기 다른 sheet로 구분되어 있다.

**5. We are assessing health effectiveness in the context of economic evaluation. There are three key outcome measures: life expectancy, quality-adjusted life years (QALYs), and disability-adjusted life years (DALYs). Imagine a life cycle with four key stages: birth, cancer incidence, death from cancer, and death without cancer (representing optimal life expectancy). In the absence of cancer, individuals experience perfect health. However, the presence of cancer can lower health-related quality of life. Use a graph (as we learned in class) to visually demonstrate how each outcome measure captures distinct aspects of health effectiveness (HINT: You do not need to calculate numbers and it is okay to show conceptually).**

텍스트, 라인, 도표, 폰트이(가) 표시된 사진

자동 생성된 설명

1. 기대수명: 출생 시 기대수명으로 시작하여 암 발병 이전까지도 기대수명은 조금씩 감소한다. 그러나 암 발병 이후에는 기대수명이 급격하게 감소하며, 암으로 인해 사망한 이후에는 0이 유지된다.
2. QALY: 출생 시 QALY는 최적의 건강 조건을 가정하여 1로 한다. 암 발병 이전에는 완벽한 건강 상태를 가정하기에 1을 유지하다가 암 발병과 함께 급격한 감소를 겪는다. 그리고 사망 이후에는 0이 된다.
3. DALY: DALY는 YLD와 YLL의 합과 같다. YLD는 장애로 인해 잃은 연수를 의미한다. 이는 암 발병 이전에는 최적의 건강 상태로 없었지만, 암 발병 이후에는 존재하게 된다. YLL은 이른 사망으로 인해 잃게 된 연수로 그래프 상에서 확인할 수 있다. DALY는 그래프 상에서 색칠된 부분의 넓이이다.

**PART B: Analytical Questions**

**6. Using the Full Year Consolidated file for 2018 and 2019, we are analyzing the association of office-based visits (including provider and physician visits) and demographic (age [AGE18X from 2018], sex [SEX], and race/ethnicity [RACETHX]) as well as socioeconomic factors (marital status [MARRY18X from 2018] health insurance coverage [INS18X from 2018], family income as a percentage of the federal poverty level [POVCAT18 from 2018], and U.S. census region [REGION18 from 2018]). Exclude individuals with missing data for any of these variables.a. Construct these two variables.**

**a. Construct the dataset.**

코드 설명:

필요한 패키지를 부착한 후 2018년과 2019년의 FYC 데이터셋을 불러와 d2018, d2019로 저장하였다. 변수들 중에서 필요한 변수만을 남겼고, OFFVST는 OBDRV18과 OBTOTV18의 합으로 정의한 d\_2018, d\_2019를 생성하였다. (2018년 기준)

**b. Start with a linear regression model. First, conduct the analysis and then interpret all results.**

코드 설명:

각 범주형 변수의 범주를 확인한 후 결측값을 제거하였다. 또한, 2018년과 2019년의 데이터를 row bind하기 위해 변수명을 통일시켰고, 합쳐진 데이터셋 pooled를 생성하였다. 뒤 문제에서는 d\_2018, d\_2019, pooled데이터셋 각각에 대한 분석을 진행하였다. 우선 선형회귀 모델인 model18\_lin, model19\_lin, model\_lin을 생성하였다.

결과:

변수가 매우 많아 2018년 데이터셋 기준 결과 해석만 작성하였다. 2019년, pooled 데이터셋도 동일한 방법으로 해석 가능하다.

Model18\_lin의 분석 결과 회귀 상수는 -1.795로 다른 모든 변수들이 0 또는 baseline일 때 OFFVST의 예상값이다. 이때 방문 횟수는 음수가 될 수 없다는 점을 참고하면 적합이 제대로 이루어지지 않았음을 예상할 수 있다. AGE 변수의 회귀계수는 0.236으로 나이가 1만큼 늘어날 때마다 반응변수의 값이 0.236 증가함을 의미한다. 또한 남성일 때에 비해 여성일 때 2.254, 인종이 1일 때에 비해 2일 때 3.108, 3일 때 0.217, 4일 때 -1.66, 5일 때 2.353 증가한다. 다음으로 결혼한 상태일 때에 비해 과부일 경우 -0.211, 이혼한 경우 1.141, 별거중인 경우 2.104, 미혼인 경우 1.924, 16세 미만인 경우 5.356 증가하였다. 16세 미만의 경우에는 inapplicable이라고 되어 있어 무시해도 되는 수치이다. 다음으로 보험이 있는 경우에 비해 있는 경우에 5.377, 빈곤한 상태에 비해 near poor는 0.591, low income은 0.879, middle income은 1.298, high income은 0.771만큼 감소하였다. 마지막으로 지역이 1인 경우에 비해 2인 경우에는 1.345, 3인 경우에는 1.085, 4인 경우에는 1.144만큼 감소하였다. 이러한 회귀계수들 중 p-value가 0.05보다 큰 RACETHX3, MARRY2, POVCAT2은 통계적으로 유의하지 않는다. 또한 R^2값이 0.08364로 모델이 반응변수의 변동을 8.36%밖에 설명하지 못함을 보여준다. F검정 결과는 통계적으로 유의하다.

Model19\_lin과 model\_lin의 회귀계수 또한 같은 방식으로 설명될 수 있다. Model19\_lin에서는 RACETHX3, POVCAT2, POVCAT5 변수의 회귀계수가 유의하지 않았고 R^2 값은 0.08606이었다. 또한, model\_lin에서는 RACETHX3, MARRY2, POVCAT2의 회귀계수가 유의하지 않았으며 R^2값은 앞선 모델들과 비슷한 값인 0.08464가 도출되었다. 이는 각 모델이 OFFVST의 변동성을 8.6%, 8.46%밖에 설명하지 못함을 의미한다.

**c. Analyze using a poisson model. Conduct the analysis and then interpret the results.**

코드 설명:

glm()과 family=’poisson’을 이용하여 포아송 모델 model18\_p, model19\_p, model\_p를 생성하였다.

결과:

우선 model18\_p의 회귀계수에서 절편인 1.1229는 모든 변수들이 0 또는 baseline일 때 OFFVST의 예상값이 exp(1.1229)임을 의미한다. 이는 0보다 크기에 선형모델의 절편보다 현실성이 있음을 보여준다. AGE의 회귀계수는 0.0209인데 이는 연령이 1 증가할 때마다 OFFVST의 로그값이 0.0209 증가함을 의미한다. 즉, 나이가 1살 늘어날 때마다 OFFVST의 예상값이 exp(0.0209)=1.02배가 됨을 알 수 있다. 다음으로 성별이 남성인 경우에 비해 여성인 경우 OFFVST 예상값은 exp(0.2422)=1.274배가 된다. 다른 변수들도 이와 같은 방법으로 회귀계수를 해석할 수 있다. 모든 회귀계수는 통계적으로 유의하다. Null deviance는 549776, residual deviance는 456543으로 두 값의 차이가 유의미하게 크다. 이를 통해 해당 모델이 null model보다 데이터를 더 잘 적합함을 알 수 있다. AIC는 542010이다. 이는 모델의 비교에 쓰일 수 있는 척도로, 낮을수록 더 나은 모델임을 의미한다.

다음으로 model19\_p에서는 MARRY2를 제외한 변수들의 회귀계수가 유의하며, null deviance 와 residual deviance의 차이가 유의미하게 크고 AIC는 518484이다. 마지막으로 pooled 데이터셋을 통해 분석한 model\_p에서는 모든 회귀계수가 유의하고 deviance의 차이도 유의미했으며 AIC는 1060952이다. Pooled 데이터셋에서 행이 훨씬 많기에 AIC가 더 큰 것으로 추정할 수 있다.

**d. Analyze using a negative binomial model. Conduct the analysis and then interpret the results.**

코드 설명:

MASS 패키지를 부착하고 glm.nb()를 통해 negative binomial model인 model18\_nb, model19\_nb, model\_nb를 생성하였다.

결과:

Model18\_nb에서 절편은 0.9845이다. 이를 통해 다른 변수들이 0이나 baseline일 때 OFFVST가 exp(0.9845)=2.678임을 알 수 있다. 또한 AGE의 회귀계수는 0.0217로, 나이가 1 증가할 때마다 OFFVST의 로그 빈도가 0.0217만큼 증가하고 예상 발생 빈도는 exp(0.0217)=1.022배가 됨을 의미한다. 그리고 남성일 때에 비해 여성일 때 예상 빈도가 exp(0.2812)=1.324배가 된다. 다른 변수들도 이와 같은 방법으로 해석할 수 있다. 회귀계수는 MARRY2, POVCAT2, POVCAT5변수에서 통계적으로 유의하지 않았다. Null deviance는 39433, residual deviance는 34154로 그 차이가 유의미하고, AIC는 185812로 포아송 모델보다 작다. 또한 dispersion parameter인 theta값은 0.525로 약간의 over dispersion을 확인할 수 있다.

Model19\_nb에서는 MARRY2, POVCAT5의 회귀계수가 유의하지 않았고 deviance차이는 유의미했으며 AIC는 174831, theta는 0.51833이었다. 마지막으로 model\_nb에서는 RACETHX3, MARRY2를 제외한 모든 변수가 통계적으로 유의했고 deviance의 차이가 컸으며 AIC는 360642, theta값은 0.52158이다.

**e. Perform a goodness-of-fit check for all models and interpret the results.**

코드 설명: 각 모델별 fit 결과를 요약하고 MAE, MSE, RMSE를 도출하였다.

결과:

2018년 데이터셋의 세개의 모델을 비교한 결과 선형모델은 fit값의 최솟값이 음수로 잘못되었다. MSE값이 선형모형에서 가장 작지만 MAE의 경우에는 음이항모델이 가장 작았으며 AIC값은 음이항 모델에서 더 작았다.

2019년 데이터셋에서는 2018년과 같이 선형모형의 fit결과의 최솟값이 음수였다. 또한 MSE값은 선형모형이, MAE값은 음이항모델이 가장 작았으며 AIC는 포아송 모델보다 음이항 모델에서 더 작았다.

Pooled 데이터셋에서도 위 두 데이터셋과 같은 결과가 도출되었다.

**f. Discuss which model is better based on the goodness-of-fit check**

MAE값이 가장 낮고 fit 결과에 음수가 포함되지 않으며, AIC값이 더 작은 negative binomial model이 데이터를 가장 잘 적합하는 것으로 판단된다.

**7. Using the Full Year Consolidated file for 2002, we are estimating health outcomes.**

**a. Develop the following health-related quality of life scores: i. Visual Analogue Scale (VAS) ii. EQ-5D iii. SF-6D iv. SF-12 physical component summary score v. SF-12 mental component summary score vi. Activities of Daily Living (ADL) vii. Instrumental Activities of Daily Living (IADL).**

코드 설명:

우선 2002년 FYC데이터셋을 d2002로 불러와 필요한 변수들만 선택한 d\_2002를 만들었다. 결측값도 모두 제거하였다. 이때 VAS는 ADHECR42변수, EQ5D는 EQU42변수, SF12\_p는 PCS42 변수, SF12\_m은 MCS42 변수를 이용하였다. SF-6D와 ADL, IADL은 아래와 같이 계산하였다.

1. SF-6D

<https://pmc.ncbi.nlm.nih.gov/articles/PMC4233005/#R11>

<https://journals.lww.com/lww-medicalcare/fulltext/2004/09000/the_estimation_of_a_preference_based_measure_of.4.aspx>

위 두 문헌 참고하여 MEPS 2002데이터셋의 변수를 이용해 SF6D를 계산하였다. 두번째 문헌은 첫번째 문헌의 주석에서 찾을 수 있었다. SF12에 대한 열두개의 응답 항목 중 두번째 문헌의 Table2를 참고해 일곱개를 골라, 두번째 문헌의 table4의 계수와 가중치를 이용하여 sf-6d를 구하였다. 일곱개의 응답 변수는 2002 FYC 데이터셋의 document를 통해 선택할 수 있었다. 변수는 PF: ADDAYA42, RL:ADPLMT42와 ADMACC42, SF:ADSOCA42,PAIN:ADPAIN42, MH: ADBLUE42, VIT:ADPEP42가 이용되었다. Table4에 명시된 가중치를 이용하여 최종적인 SF-6D를 계산할 수 있었다.

텍스트, 스크린샷, 폰트, 문서이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 폰트, 문서이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 번호, 메뉴이(가) 표시된 사진

자동 생성된 설명

1. ADL, IADL:

ADL과 IADL 항목 계산에는 변수 중 ADLHLP31, ADLHLP42, ADLHLP53, ADL3MO31, ADL3MO42, ADL3MO53, IADLHP31, IADLHP42, IADLHP53, IADL3M31, IADL3M42, IADL3M53이 이용되었다. ADL은 ADLHLP31, ADLHLP42, ADLHLP53, ADL3MO31, ADL3MO42, ADL3MO53의 합으로, 코드북을 참고한 바에 따르면 yes가 1, no가 2로 입력되어 있었기에 2를 0으로 바꿔서 최종 계산에 포함시켰다. 또한, ADL3MO, IADL3M이 포함된 변수들은 결측값이 매우 많아 결과를 혼동시키기에 결측값을 0으로 바꿔 계산하였다. 그에 따라 adl\_score, iadl\_score 변수는 최솟값이 0, 최댓값이 6인 칼럼이 되었다. 마지막에 변수별 결측값의 개수를 확인하였고, VAS, EQ5D, sf6d\_score, SF12\_p, SF12\_m, adl\_score, iadl\_score만을 포함한 데이터셋 ‘final’를 생성하였다.

**b. Calculate the mean and standard deviation for each measure.**

코드 설명:

Mean(), sd()함수를 이용하여 각 score 별 평균과 표준편차를 구하였다.

결과:

평균- VAS 2.770152, EQ5D 0.0108707, SF6D 0.5447209,SF12\_p 29.49, SF12\_m 30.84635, ADL 0.09110012, IADL 0.163428

표준편차: VAS 4.801735, EQ5D 1.331881, SF6D 0.06832193 ,SF12\_p 25.54017, SF12\_m 26.36463, ADL 0.6269851, IADL 0.8337113

**c. Analyze the correlation for each measure using the cor( ) function in R**

‘김연주\_p2\_분석결과.pdf’ 참고

**d. Select two measures from the list above and conduct a regression analysis to examine their associations with age, sex, race/ethnicity, marital status, family income as a percentage of the federal poverty level, and U.S. census region. In your analysis, specify the regression model you chose and explain your reasoning for selecting this model.**

코드 설명: 포아송 모델과 선형 회귀 모델을 생성하고 mae, mse, rmse를 측정하였다.

결과:

결측값이 없는 VAS와 EQ5D를 분석하였다. 우선 예측변수에 존재하는 결측값을 모두 제거하고 회귀 모델을 만들었다.

1. VAS: VAS 값은 그 결과가 0부터 10까지의 정수이기 때문에 연속형 변수로 볼 수 있지만 분석 결과 선형회귀 모델보다 예측력이 높았던 포아송 모델을 사용하였다..
2. EQ5D: 연속형 변수이기 때문에 선형회귀 모델을 사용하였다.
3. Mae, mse, rmse 모두 두 모델에서 낮은 값을 보였다.

**e. Interpret the results for the two measures and compare whether the associations with each variable are consistent.**

1. VAS: 분석 결과 vas\_model의 회귀계수에서 절편인 2.0207는 모든 변수들이 0 또는 baseline일 때 VAS의 예상값이 exp(2.0207)임을 의미한다. 이는 0보다 크기에 데이터가 잘 적합되었음을 보여준다. AGE의 회귀계수는 0.0016인데 이는 연령이 1 증가할 때마다 VAS의 로그값이 0.0016 증가함을 의미한다. 즉, 나이가 1살 늘어날 때마다 VAS의 예상값이 exp(0.0016)배가 됨을 알 수 있다. 다음으로 성별이 남성인 경우에 비해 여성인 경우 VAS 예상값은 exp(0.013)배가 된다. 다른 변수들도 이와 같은 방법으로 회귀계수를 해석할 수 있다. RACETHNX2, RACETHNX3, RACETHNX4, MARRY2, MARRY4, MARRY5, REGION2를 제외한 모든 회귀계수는 통계적으로 유의하다. Null deviance는 8563, residual deviance는 8388.8으로 두 값의 차이가 유의미하게 크다. 이를 통해 해당 모델이 null model보다 데이터를 더 잘 적합함을 알 수 있다. AIC는 73571이다.
2. EQ5D: Eq5d\_model의 분석 결과 회귀 상수는 0.9679로 다른 모든 변수들이 0 또는 baseline일 때 EQ5D의 예상값이다. 이때 EQ5D 값은 -0.6에서 1.0이라는 사실을 고려하면 적합이 잘 이루어졌음을 알 수 있다. AGE 변수의 회귀계수는 -0.0044으로 나이가 1만큼 늘어날 때마다 반응변수의 값이 0.0044 감소함을 의미한다. 또한 남성일 때에 비해 여성일 때 0.0232 감소한다. 다른 변수들도 이와 같은 방법으로 회귀계수를 해석할 수 있다 이러한 회귀계수들 중 p-value가 0.05보다 큰 RACETHNX2, MARRY2, REGION2, REGION4은 통계적으로 유의하지 않는다. 또한 R^2값이 0.1436으로 모델이 반응변수의 변동을 14.36% 설명함을 보여준다. F검정 결과는 통계적으로 유의하다.
3. 두 measurement에 모두 유의미한 영향을 미친 변수는 성별과 연령이다. 그러나 연령의 경우 두 모델에서 영향의 방향이 상반되었다. 소득수준은 EQ5D에서만 유의미한 영향을 보였다. 두 모델에서 일부 변수들은 일관된 방향성을 보였으나 연령과 소득 중 특정 변수는 각 지표에서 상반된 영향을 미쳤다. 이를 통해 VAS와 EQ5D가 각기 다른 건강 관련 요소를 반영할 가능성이 있음을 알 수 있다.