**Data Science in Healthcare**

**Problem Set #1**

이름: 김연주

학과: 통계학과

학번: 2021250461

**PART A: Conceptual Questions**

**1. We are analyzing health care spending.**

**a. Explain what the following statistical measures represent: mean, variance, kurtosis, and skewness.**

(1) mean (평균): 모든 자료의 값을 더하여 자료의 수로 나눈 값이다. 이는 극단적으로 아주 작거나 큰 값이 포함된 경우 영향을 많이 받기에 극단값이 존재하는 경우 중위수 등 다른 대푯값을 이용하기도 한다.

(2) variance (분산): 각 데이터에서 평균을 뺀 값을 제곱한 것들의 평균이다. 데이터의 관측값들이 평균으로부터 퍼져 있는 정도를 나타낸다. 분산을 통해 자료가 평균에서 퍼진 정도와 분포를 유추할 수 있다.

(3) kurtosis (첨도): 자료 분포의 뾰족한 정도를 나타낸다. 3보다 큰 첨도는 일반적으로 분포가 정규분포보다 더 뾰족하고 꼬리가 얇음을 의미하고, 3보다 작은 첨도는 자료의 분포가 정규분포보다 납작하고 꼬리가 두꺼운 경우를 의미한다. 즉, 자료가 정규분포를 따를 때 첨도는 3이 된다.

(4) skewness (왜도): 왜도가 0이라면 데이터의 분포가 대칭적임을 의미한다. 왜도가 양수라면 오른쪽으로 치우친 분포, 왜도가 음수라면 왼쪽으로 치우친 분포를 보인다.

**b. Suppose our measure of healthcare spending follows a normal distribution. What would the kurtosis and skewness be?**

주어진 자료의 분포가 정규분포를 따른다면 첨도는 3, 왜도는 0이 된다. 첨도는 자료가 정규분포일 때 3의 값을 갖고, 정규분포는 평균을 기준으로 좌우 대칭이기 때문이다.

**2. We are considering using a linear regression model to examine the association between health care spending and age.**

**a. Explain how this model is developed and under what conditions it is most appropriate to use. At this stage, there is no need to discuss specific assumptions.**

의료비지출과 연령 사이의 관계를 선형회귀 모형을 통해 알아볼 수 있다. 독립변수를 연령, 종속변수를 의료비 지출로 한 단순선형회귀 분석이 가능하다. 이러한 선형회귀모형은 독립변수와 종속변수가 선형 관계를 지녀야 적절하게 이용될 수 있다.

Y=β0​+β1​X+ϵ 형태의 모델이며 오차항인 ϵ를 최소화하는 모델을 사용하는 것이 가장 적합하다.

**b. What assumptions must be satisfied in order to use a linear regression model? Explain each assumption.**

(1) 선형성 가정: 독립변수와 종속변수의 관계가 선형 관계이어야 한다.

(2) 내생성의 부재: 독립변수와 오차항이 무관해야 한다. 즉, 위 모형에서 X와 ϵ의 상관계수가 0이어야 한다. 이는 회귀계수의 추정치가 편향되지 않도록 한다. 관찰되지 않은 confounder가 존재하면 충족하기 어렵다.

(3) 정규성 가정: 오차항이 무작위로 발생해야 하며 정규 분포를 따라야 한다.

(4) 독립성 가정: 오차항은 서로 독립이어야 한다.

(5) 등분산성 가정: 오차항들의 분산은 독립변수 값에 관계없이 모두 동일해야 한다.

**c. Discuss whether these assumptions are likely to be satisfied when examining the association between health care spending and age in real-world settings.**

(1) 선형성 가정: 현실 세계에서 의료비 지출과 연령의 관계가 항상 선형적일 것이라는 보장은 없다. 젋은 층에서는 의료비 지출이 거의 없다가 고령층에서 급격히 증가하는 양상을 보일 수 있기 때문이다.

(2) 내생성의 부재: 현실 세계에서 의료비 지출과 연령 모두에 영향을 주는 confounder가 존재한다. 대표적인 예로 건강 상태가 있다. 그렇기에 내생성이 존재할 가능성이 높다.

(3) 정규성 가정: 의료비 지출은 일부 사람들에게서 극단적으로 높게 나올 수 있고 비대칭적일 가능성이 높다. 그렇기에 오차항이 정규 분포를 따르지 않을 가능성이 크다.

(4) 독립성 가정: 특정 질병을 공유하거나 같은 가구에 속한 사람들은 비슷한 의료비 지출 패턴을 보일 가능성이 있기에 독립성 가정을 하기 어렵다.

(5) 등분산성 가정: 높은 연령대일수록 소득이나 건강 상태 등으로 인해 사람들 간의 의료비 지출의 변동성이 커질 가능성이 있다. 즉, 연령이 높아짐에 따라 오차항의 분산이 커지는 것이다. 이는 등분산성 가정이 어려움을 시사한다.

**3. We are examining the association between health care spending and mental health status with hypothesis testing.**

**a. State the null hypothesis.**

귀무가설은 의료비 지출과 정신건강 상태가 서로 무관하다는 것이다. 즉, 독립변수를 정신 건강 상태, 종속변수를 의료비 지출이라고 하고 세운 단순선형회귀 모형의 회귀 계수가 0임을 의미한다.

**b. State the alternative hypothesis**

대립가설은 의료비 지출과 정신건강 상태가 무관하지 않고 서로 관련이 있다는 것이다. 즉 회귀계수가 0이 아님 (양측검정에서)을 의미한다.

**c. To reject the null hypothesis, which measure should be examined? Explain what this measure indicates.**

귀무가설을 기각하기 위해서는 p-value가 필요하다. P-value는 귀무가설이 사실이라는 가정 하에서 관찰된 값 (또는 같거나 더 극단적인 값)을 얻을 확률을 의미한다. 즉, 해당 상황에서는 의료비 지출과 정신건강 상태가 실제로는 무관한 상태에서 자료와 같은 관찰값을 얻을 확률을 의미하는 것이다. 그러므로 p-value가 작을수록 귀무가설이 참일 확률이 낮음을 의미하고 귀무가설 기각의 근거가 된다. 일반적으로 유의수준 0.05보다 p-value가 작으면 귀무가설을 기각한다.

**4. After developing a model for health care spending, it is necessary to evaluate its performance. We learned about two evaluation metrics in class. Explain these two metrics and what they measure.**

(1) : 결정계수는 회귀모형 속 독립변수로 인해 설명될 수 있는 종속변수의 비율을 의미한다. 즉, 모델이 종속 변수의 변동성을 얼마나 잘 설명하는지를 나타내는 지표인 것이다. 이는 0부터 1까지의 값을 갖고 1에 가까울수록 종속변수를 더욱 잘 설명하는 모델임을 의미한다.

(2) Predictive Ratio: 모델로 인해 생성된 예측값과 실제값의 비율로 모델의 정확성을 평가한다. 이는 예측된 값을 실제 값으로 나눈 지표이고 1에 가까울수록 모델의 예측이 정확함을 의미한다. 또한 1보다 큰 경우에는 예측이 과대평가된 것이고 1보다 작은 경우에는 예측이 과소평가된 것이므로 예측의 오차 방향을 확인하기에도 용이하다.

**PART B: Analytical Questions**

**5. We are analyzing two measures of health care spending from 2019 to 2022: annual total health care spending and annual out-of-pocket spending. To measure spending, we are focusing on the following settings: hospital inpatient stays, outpatient visits, office-based medical provider visits, emergency room visits, and prescribed medications.**

**a. Construct these two variables.**

코드 설명:

1. 데이터 로드: 2019년부터 2022년까지의 FYC 데이터셋을 불러와 ‘d연도’로 저장한다.
2. 변수 선택: 데이터셋의 변수명을 확인한 후 DUPERSID, VARSTR, VARPSU, total expenditure, out-of-pocket expenditure, PERWT의 변수로 이루어진 새로운 데이터셋을 만든다. 이때 주제별로 total expenditure과 out-of-pocket expenditure 변수가 상이하여 코드북을 참고하였고 각 주제별 exp변수를 더해 ‘total\_exp’와 ‘pocket\_exp’ 변수를 생성하였다. 또한, year 변수를 추가하여 추후 inflation 계산에 용이한 형태로 만들었다.
3. Merge: 네개의 데이터셋을 row bind하여 pooled라는 데이터셋을 형성하였다. 이때 가중치는 데이터셋의 개수인 4로 나누어 poolwt 변수를 추가했다.

결과: 107084개의 행과 55개의 열로 이루어진 pooled 데이터셋이 생성되었다.

**b. Estimate the average values for these two variables using survey weights and inflation adjustments.**

코드설명:

1. Inflation 고려: 강의자료의 CPI를 참고하여 연도별 CPI를 ‘inflation\_factors’에 저장하였다.
2. 2022년 기준 물가 조정 비율 계산: 2022년을 base year로 하여 연도별 물가 조정 비율을 구하였다. 이후 2022년을 기준으로 값이 조정된 ‘total\_exp\_adj’와 ‘pocket\_exp\_adj’ 변수를 생성하였다.
3. Define the survey design: 표본 단위, 층화, 가중치를 고려한 ‘dsgn’을 생성했다.
4. Average: 가중치와 inflation을 고려한 평균은 total expenditure에서 6062.7$, out-of-pocket expenditure에서 613.2$의 값을 보였다.

결과: total expenditure 6062.7$, out-of-pocket expenditure 613.2$

**c. Estimate the average values at the 25th, 50th, and 75th percentiles for these two variables using survey weights and inflation adjustments.**

코드 설명:

25th, 50th, 75th quantile: svyqualtile()을 이용하여 구하였다.

결과: total spending에서 q1 105$, q2 843.5013$, q3 4071.7030값이 도출되었고, out-of-pocket spending에서는 q1 0$, q2 62$, q3 418.3839$값이 도출되었다.

**d. Draw a histogram for each variable and then state whether these variables follow a normal distribution.**

코드 설명:

1. Total\_exp\_adj 히스토그램: pooled 데이터셋의 total\_exp\_adj의 히스토그램을 그렸다.
2. 로그변환 histogram: 변수 값의 범위가 매우 크고 편향되어 일반 히스토그램으로는 분포 파악이 힘들었다. 따라서 total\_exp\_adj를 로그 변환하여 극단적인 값을 보정하는 효과를 기대하였다. 또한 이때 0인 값도 로그 변환이 가능하도록 하기 위해 아주 작은 값인 0.0001을 더하여 로그변환하였다.
3. Pocket\_exp\_adj 히스토그램: total\_exp\_adj와 같은 과정
4. 로그변환 histogram: total\_exp\_adj와 같은 과정

결과: 일반 히스토그램에서는 두 변수 모두 분포 파악이 불가능하였다. 로그 변환 히스토그램을 통해 두 변수 모두 작은 값에 빈도가 몰려 있다가 그 이후 정규분포와 유사한 형태를 가짐을 알 수 있었다.

**6. Using the Full Year Consolidated file for 2020 and 2022, we are analyzing the association of health care spending and demographic (age, sex, and race/ethnicity) as well as health factors (diagnoses of cancer, stroke, and diabetes). Exclude individuals with missing data on these variables.**

**a. Construct the dataset.**

코드 설명:

1. 데이터 로드: 2020년과 2022년의 FYC데이터를 불러와 각각 ‘d연도’로 저장하였다. 이후 관심 변수인 age (AGE연도X), sex (SEX), race (RACETHX), diagnosis of cancer (CANCERDX), diagnosis od stroke (STRKDX), diagnosis of diabetes (DIABDX\_M18)만을 남긴 ‘d\_연도’ 데이터셋을 만들었다.
2. 결측치 제거: sum(is.na())를 통해 d\_2020과 d\_2022의 결측치 개수를 확인한 결과 모두 0개였다. 이는 수치형 변수에 결측값이 존재하지 않고 범주형 변수의 경우 결측치가 다른 숫자로 기입되어 있을 가능성을 시사한다. Table()을 통해 범주형 변수의 범주를 살펴보고 코드북을 참고한 결과 1과 2를 제외한 값은 결측값임을 알 수 있었다. 그래서 1과 2가 아닌 값은 모두 NA로 변환하고 na.omit()을 통해 결측값을 제거하였다.

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

코드 설명: ‘TOTEXP’를 반응변수로 하는 다중선형회귀 모형 model20\_1과 model22\_1을 만들었다. 범주형 변수는 factor()를 사용하여 가변수를 생성하였다.

결과:

1. model20\_1: 잔차의 분포를 통해 실제값과 예측된 값의 차이가 꽤 큼을 알 수 있다. 회귀계수를 보면 나이가 1단위 증가할 때마다 반응변수가 68.722증가하고 성별이 2인 경우에 성별이 1인 경우보다 반응변수가 825.237 더 많다. 또한 RACETHX가 1인경우에 비해 2인 경우에는 1837.124, 3인 경우에는 1701.786, 5인 경우에는 3927.34 증가하고 4인 경우에는 1087.107 감소한다. Cancer, stroke, diabetes의 진단을 받지 않은 경우에는 받은 경우보다 각각 8419.361, 7194.56, 6713.584 적게 지출함을 알 수 있다. 이러한 회귀계수는 RACETHX4를 제외하고 모두 유의하며 F통계량이 크고 p-value가 작아 모델이 통계적으로 유의하다. 또한 R^2값이 0.04769로 해당 모델이 반응변수의 변동성을 4.769% 설명함을 알 수 있다.
2. model22\_1: 잔차의 분포를 통해 실제값과 예측된 값의 차이가 꽤 큼을 알 수 있다. 회귀계수를 보면 나이가 1단위 증가할 때마다 반응변수가 59.55증가하고 성별이 여자인 경우에 남성보다 반응변수가 1653.56 더 많다. 또한 RACETHX가 1인경우에 비해 2인 경우에는 3714.18, 3인 경우에는 1258.18, 4인 경우에는 236.26, 5인 경우에는 3505.2 증가한다. Cancer, stroke, diabetes의 진단을 받지 않은 경우에는 받은 경우보다 각각 7394.16$, 6939.55$, 8093.33$ 적게 지출함을 알 수 있다. 이러한 회귀계수는 RACETHX3과 RACETHX4를 제외하고 모두 유의하며 F통계량이 크고 p-value가 작아 모델이 통계적으로 유의하다. 또한 R^2값이 0.04854로 해당 모델이 반응변수의 변동성을 4.854% 설명함을 알 수 있다.

**c. Perform a goodness-of-fit check for this model and interpret the results.**

코드 설명: 예측값에 대한 summary()를 통해 예측된 값의 특징을 파악하였다. 이후 R^2값과 mae, mse, rmse값을 구했다.

결과:

1. model20\_1: 예측값의 최솟값이 -1255인데 의료비 지출은 음수가 될 수 없으므로 잘못된 예측을 하고 있음을 알 수 있다. R^2값이 0.04768774로 매우 낮으며 이는 해당 모델이 반응변수의 변동을 약 4.77%밖에 설명하지 못함을 의미한다. MAE는 8368.911으로 예측값에서 실제값을 뺀 절댓값의 평균이 매우 크다는 것을 알려준다. MSE (=555210614)와 RMSE (23562.91)도 매우 큰 값을 보인다. 이를 통해 해당 모델이 반응변수를 설명하기에는 설명력이 부족하다는 것을 알 수 있다.
2. model22\_1: 예측값의 범위는 크게 상식에서 벗어나지 않고 있다. 그러나 R^2값이 0.0485441로 매우 낮으며 이는 해당 모델이 반응변수의 변동을 약 4.85%밖에 설명하지 못함을 의미한다. MAE는 9469.516으로 예측값에서 실제값을 뺀 절댓값의 평균이 매우 크다는 것을 알려준다. MSE (=581276588)와 RMSE (24109.68)도 매우 큰 값을 보인다. 이를 통해 해당 모델도 반응변수를 설명하기에는 설명력이 부족하다는 것을 알 수 있다.

**d. Consider an alternative model. Conduct the analysis and then interpret all results.**

코드 설명:

1. Log-transformed model: TOTEXP에 0.00001을 더해 로그 변환한 값을 새로운 변수 ‘TOTEXP\_1’에 저장하였다. 0인 값도 같이 분석하기 위해 아주 작은 값인 0.00001을 더하였다. 이후 TOTEXP\_1을 반응변수로 하는 모델 ‘model20\_2’와 ‘model22\_2’를 생성하였다.
2. glm: 의료비가 0보다 큰 경우에 대하여 generalized linear model ‘model20\_3’과 ‘model22\_3’을 만들었다. 이후 감마 분포를 기반으로 로그 링크를 이용한 generalized linear model ‘model20\_4’와 ‘model22\_4’를 만들었다.
3. Two-part model: 의료비 지출 데이터는 지출이 발생하지 않은 0인 관측값이 존재한다. 이에 따라 모델을 두가지로 나누어서 첫번째 파트에서는 의료비의 발생 여부를 예측하고, 두번째 파트에서는 의료비가 발생했을 때 그 금액을 예측한다. 그에 따라 part1에서는 로짓 링크를 통해 로지스틱 회귀 분석을 실행했고, part2에서는 감마 분포와 로그 링크를 사용했다. 이러한 ‘model20\_5’와 ‘model22\_5’를 생성하였다.

결과:

1. model20\_2, model22\_2: 우선 두 모델에서 잔차의 범위가 넓기에 모델의 예측이 실제값과 다름을 알 수 있다. Model20\_1의 회귀계수를 통해 나이가 1살 증가할 때마다 총 의료비 지출의 로그 변환값이 0.0793 증가함을 알 수 있다. 성별이 여성일 때 남성에 비해 의료비 지출의 변환값이 1.882 증가하고 인종이 1인 경우에 비해 2인 경우에는 3.736, 3인 경우에는 1.234, 4는 1.759, 5는 2.433 증가한다. Cancer, stroke, diabetes의 진단을 받지 않은 경우에는 받은 경우보다 총의료비 지출의 로그값이 각각 1.025, 1.67, 2.79 적음을 알 수 있다. 그리고 model22\_2에서 나이가 1살 증가할 때마다 총 의료비 지출의 로그 변환값이 0.0806 증가함을 알 수 있다. 또한 성별이 여성일 때 남성에 비해 의료비 지출의 변환값이 1.567 증가하고 인종이 1인 경우에 비해 2인 경우에는 4.228, 3인 경우에는 1.925, 4는 2.822, 5는 3.213 증가한다. Cancer, stroke, diabetes의 진단을 받지 않은 경우에는 받은 경우보다 총의료비 지출의 로그값이 각각 1.082, 1.125, 2.587 적음을 알 수 있다. 이러한 회귀계수는 모두 유의하며 F통계량이 크고 p-value가 작아 모델이 통계적으로 유의하다. 또한 R^2값이 각각 0.176, 0.186으로 해당 모델들이 반응변수의 변동성을 17.6%, 18.6% 설명함을 알 수 있다.
2. model20\_3, model22\_3: model20\_3에서 회귀계수를 통해 나이가 한 단위 증가할 때마다 총 의료비 지출이 53.54 증가함을 알 수 있다. 또한 성별이 여성일 때 남성에 비해 의료비 지출이 409.25 증가하고 인종이 1인 경우에 비해 2인 경우에는 1074.61, 3인 경우에는 1826.09, 5는 4255.44 증가하고 4의 경우에는 1941.73 감소한다. Cancer, stroke, diabetes의 진단을 받지 않은 경우에는 받은 경우보다 총의료비 지출이 각각 8621.31, 6972.11, 6325.1 적음을 알 수 있다. 이러한 회귀계수는 성별 변수를 제외하고는 모두 통계적으로 유의하다. 또한 null deviance와 residual deviance의 차이가 488731149518로 매우 크므로 LRT결과가 유의할 것이라 예측할 수 있다. 그리고 model22\_3에서 회귀계수를 통해 나이가 한 단위 증가할 때마다 총 의료비 지출이 38.78 증가함을 알 수 있다. 또한 성별이 여성일 때 남성에 비해 의료비 지출이 1436.1 증가하고 인종이 1인 경우에 비해 2인 경우에는 3005.46, 3인 경우에는 1021.69, 5는 3285.36 증가하고 4의 경우에는 563.01 감소한다. Cancer, stroke, diabetes의 진단을 받지 않은 경우에는 받은 경우보다 총의료비 지출이 각각 7332.4, 6898.1, 7770.76 적음을 알 수 있다. 이러한 회귀계수는 RACETHX 3, RACETHX 4를 제외하고는 모두 통계적으로 유의하다. 또한 null deviance와 residual deviance의 차이가 399684545274로 매우 크므로 LRT결과가 유의할 것이라 예측할 수 있다.
3. model20\_4, model22\_4: model20\_4에서 나이가 한 단위 증가할 때마다 총 의료비 지출이 0.88% 증가함을 알 수 있다. 또한 성별이 여성일 때 남성에 비해 의료비 지출이 12.15% 증가하고 인종이 1인 경우에 비해 2인 경우에는 20.8%, 3인 경우에는 22.14%, 5는 45.47% 증가하고 4의 경우에는 29.3% 감소한다. Cancer, stroke, diabetes의 진단을 받지 않은 경우에는 받은 경우보다 총의료비 지출이 각각 78.1%, 56.39%, 66.4% 적음을 알 수 있다. 이러한 회귀계수는 모두 통계적으로 유의하고 null deviance와 residual deviance의 차이가 5375.473로 매우 크므로 자유도에 관계 없이 LRT결과가 유의할 것이라 예측할 수 있다. 또한, model22\_4에서 나이가 한 단위 증가할 때마다 총 의료비 지출이 0.45% 증가함을 알 수 있다. 또한 성별이 여성일 때 남성에 비해 의료비 지출이 19.87% 증가하고 인종이 1인 경우에 비해 2인 경우에는 31.45%, 3인 경우에는 10.51%, 5는 33.55% 증가하고 4의 경우에는 11.56% 감소한다. Cancer, stroke, diabetes의 진단을 받지 않은 경우에는 받은 경우보다 총의료비 지출이 각각 59.18%, 50.8%, 65.16% 적음을 알 수 있다. 이러한 회귀계수는 t-test 결과 RACETHX 3,4,5를 제외하고 모두 통계적으로 유의하다. 또한 null deviance와 residual deviance의 차이가 3185.545로 매우 크므로 LRT결과가 유의할 것이라 예측할 수 있다.
4. model20\_5, model22\_5: 우선 model20\_5의 part1에서 회귀계수를 통해 나이가 1살 증가할 때마다 총 의료비 지출이 발생하는 로그 오즈의 값이 0.029047 증가함을 알 수 있다. 오즈는 사건이 발생하는 확률을 발생하지 않을 확률로 나눈 것으로, 이 상황에서는 의료비 지출이 발생할 확률을 지출이 0일 확률로 나눈 값이다. 로그 오즈는 해당 오즈에 로그를 취한 값이다. 성별이 여성일 때 남성에 비해 의료비 지출의 로그 오즈는 0.1214 증가하고 인종이 1인 경우에 비해 2인 경우에는 0.208, 3인 경우에는 0.2214, 5는 0.4547 증가하며 4인 경우에는 0.293 감소한다. Cancer, stroke, diabetes의 진단을 받지 않은 경우에는 받은 경우보다 총의료비 지출의 로그 오즈값이 각각 0.781, 0.564, 0.664 적음을 알 수 있다. Part2는 감마분포를 사용한 모델이다. 나이가 1살 증가할 때마다 총 의료비 지출의 로그 변환값이 0.0088 증가함을 알 수 있다. 또한 성별이 여성일 때 남성에 비해 의료비 지출의 변환값이 0.1215 증가하고 (여성일 경우 의료비 지출이 12.15% 더 많음) 인종이 1인 경우에 비해 2인 경우에는 0.208, 3인 경우에는 0.221, 5는 0.455 증가하며 4는 0.293 감소한다. Cancer, stroke, diabetes의 진단을 받지 않은 경우에는 받은 경우보다 총의료비 지출의 로그값이 각각 0.781, 0.564, 0.664 적음을 알 수 있다. 그리고 model22\_5의 회귀계수도 같은 방법으로 해석할 수 있다. 모든 모델에서 deviance값의 차는 크기 때문에 LRT결과도 통계적으로 유의할 적이라 예측 가능하다.

**e. Perform a goodness-of-fit check for this model and interpret the results.**

1. model20\_2, model22\_2: 우선 model20\_2에서는 예측값의 최솟값이 -3.499인데 의료비 지출은 음수가 될 수 없으므로 잘못된 예측을 하고 있음을 알 수 있다. R^2값이 0.176049로 매우 낮으며 이는 해당 모델이 반응변수의 변동을 약 17.6%밖에 설명하지 못함을 의미한다. MAE는 7439.352으로 예측값에서 실제값을 뺀 절댓값의 평균이 매우 크다는 것을 알려준다. MSE (=638320284)와 RMSE (25265)도 매우 큰 값을 보인다. 이를 통해 해당 모델이 반응변수를 설명하기에는 설명력이 부족하다는 것을 알 수 있다. 다음으로 model22\_2는 예측값의 최솟값이 -3.414인데 의료비 지출은 음수가 될 수 없으므로 이 모델 또한 잘못된 예측을 하고 있음을 알 수 있다. R^2값이 0.1859659로 매우 낮으며 이는 해당 모델이 반응변수의 변동을 약 18.6%밖에 설명하지 못함을 의미한다. MAE는 8897.836으로 예측값에서 실제값을 뺀 절댓값의 평균이 매우 크다는 것을 알려준다. MSE (=690063860)와 RMSE (26269.07)도 매우 큰 값을 보인다. 이를 통해 해당 모델이 반응변수를 설명하기에는 설명력이 부족하다는 것을 알 수 있다.
2. Glm 모델들의 적합도 평가: molde20\_3/4/5, molde22\_3/4/5는 glm 모델이기 때문에 deviance와 자유도를 이용한 카이제곱 검정을 실행하였다. 우선 예측변수를 모두 제외하고 상수항만 남은 모델 null\_model1과 null\_model2를 생성하였다. 이제 null 모델과 model20\_3을 비교해보면 상수항만 남은 모델보다 직접 만든 모델의 적합도가 더 나은지에 관한 평가가 가능하다. 카이제곱 검정 시 자유도는 두 모델의 parameter 개수 차이를 나타낸다. 예시로 model\_20\_3의 경우 가변수까지 모두 포함한 모수의 개수가 9개이고 null model은 0개이기 때문에 자유도는 9가 된다. Null model과 비교하려는 model의 deviance 차이값과 자유도를 통해 카이제곱 검정을 하여 그 결과가 유의하면 (p-value가 0.05보다 작으면) 직접 만든 모델의 적합도가 더 나음을 의미한다. 검정 결과 모든 glm모델들의 적합도가 유의함을 알 수 있었다. Model20\_5, model22\_5의 경우 코드로는 나타나지 않았지만 deviance값의 차이가 매우 커서 카이제곱 검정 결과가 유의함을 유추할 수 있었다.

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

1. Model20\_2, model22\_2: 설명력이 부족하여 비교에서 제외하였다.
2. Model3 vs model 4: model20(22)\_3과 model20(22)\_4 또한 deviance를 통해 비교 가능하다. 잔차 분산 (residual deviance)는 잔차 제곱합을 자유도로 나눈 값으로, 이는 모델이 예측한 값과 실제 값의 차이가 얼마나 퍼져 있는지를 나타내는 지표이다. 즉, 잔차분산이 크면 모델이 데이터를 잘 설명하지 못함을 의미하는 것이다. 위의 (e)에서 null model에 비해 설명력이 높다는 결과가 있었다고 해도 다른 모델에 비해 residual deviance가 유의미하게 차이 난다면 설명력이 부족할 수 있는 것이다. 그에 따라 잔차 분산을 비교해본 결과 model20\_4와 model22\_4의 잔차분산이 훨씬 작아 더욱 설명력이 높았다. 또한, AIC (Akaike Information Criterion)으로도 두 모델을 비교할 수 있다. AIC는 모델의 설명력을 평가하는 동시에 과적합 방지를 위한 기능도 있기에 여러 모델을 비교하는 데에 있어서 좋은 지표가 된다. 이는 모델의 적합도가 높을수록, 불필요한 복잡도가 낮을수록 그 값이 작아진다. 해당 모델에서는 model20(22)\_4의 AIC가 더 낮으므로 데이터를 비교적 잘 설명한다고 할 수 있다.
3. Model4 vs model5: model20(22)\_4와 model20(22)\_5의 비교 결과는 다음과 같다. Model20(22)\_5는 two part 모델로 의료비가 0인 경우를 함께 분석한다. 해당 데이터는 0인 값이 많이 존재하기에 model4처럼 0보다 큰 값만을 분석한다면 중요한 정보를 놓쳐 설명력이 하락할 수 있다. Two part 모델은 의료비 지출의 이분법적 특성 (0 또는 0 초과)를 반영하기에 더 높은 적합도를 지닐 것이라 기대할 수 있다. 그렇기에 AIC와 잔차분산이 동일하여도 의료비 지출이 0인 경우를 같이 고려하는 two part model이 더 낫다고 평가할 수 있다.

**g. Focus on the association between dementia and healthcare spending. Consider interpreting these results as causal effects, specifically the impact of dementia on healthcare spending. Discuss whether this interpretation is accepted. If it is, provide supporting reasons; if not, outline the reasons for the rejection.**

코드 설명:

1. 데이터 로드: 치매 진단과 관련된 직접적인 변수가 없기에 인지 능력을 나타내는 ‘COGLIM31’과 ‘DFCOG42’를 이용하였다.
2. 상관관계: cor()함수를 통해 의료비와 인지능력의 상관관계를 알아보았다. 대부분 -0.1~-0.2의 값을 보였다. 이는 인지능력의 저하가 존재할 때 (COGLIM==1) 의료비 지출이 줄어듦을 의미하긴 하지만 그 차이가 크지 않다. 즉, 선형적인 상관관계가 강하지 않다.
3. 회귀분석: 분석 결과 개별 회귀계수와 F통계량은 유의했지만 R^2값이 2020년 자료에는 0.03498, 2022년 자료에서는 0.02202로 모델이 의료비 지출의 변동성을 약 2~3%밖에 설명하지 못한다.

결과: 치매와 의료비 지출 사이에는 상관관계가 존재한다고 할 수 있지만 인과관계는 그 존재를 단언할 수 없다. 치매 환자는 다양한 다른 질환들을 동반할 가능성이 높아 의료 서비스 이용이 증가한다. 특히 치매의 주요 증상 중 하나인 인지 기능의 저하는 장기 요양 서비스를 필요로 하는데 이는 많은 의료비 지출을 초래한다. 그러나 이 둘이 인과관계를 가진다고는 할 수 없다. 그 이유는 다음과 같다.

1. 혼란 변수의 존재 가능성: 치매와 의료비 지출의 관계 사이에는 수많은 confounder가 존재한다. 예를 들어 나이는 치매 발병 여부와 의료비 지출 모두에 영향을 주는 대표적인 혼란변수이다. 나이뿐만 아니라 건강행태, 사회경제적 상태, 동반 질환 등도 혼란변수로 작용할 수 있다.
2. 역 인과관계의 가능성: 치매가 의료비 지출을 증가시킬 수 있지만, 반대로 높은 의료비 지출이 치매의 진단이나 관리에 영향을 미칠 수도 있다. 예를 들어 평소 건강상태에 관심이 많아 의료비 지출이 높은 사람들에게서 치매 진단율이 높을 가능성이 있는 것이다.

이러한 이유로 인해 치매와 의료비 지출 간의 직접적 인과관계는 단언지을 수 없다. 또한, R 코드 결과를 참고해보면 인지능력과 의료비 지출 사이에 선형적인 상관관계도 분명하지 않고, 회귀 분석 결과도 인과관계의 부재를 뒷받침한다. 선형회귀 분석 결과에서 각각의 회귀계수와 F통계량은 유의하지만 R^2값이 매우 낮다. 이는 추가적인 변수가 필요함을 시사하며 나아가 혼란변수의 존재를 의미하기도 한다. 그렇기에 치매와 의료비 지출의 인과관계는 인정될 수 없다.