**Data Science in Healthcare**

**Problem Set #4**

이름: 김연주

학과: 통계학과

학번: 2021250461

**PART A: Analytical Questions.**

**Use the MEPS HC-243: 2022 Full Year Consolidated Data File for this analysis. From the dataset, identify and generate the following 14 variables. Each variable must be created as a new variable based on the corresponding original variable described in parentheses.**

**•bmi (ADBMI42): body mass index.**

**• age (AGE22X): age in years.**

**• gender (SEX): equal to 1 if male.**

**• race levels (RACETHX): 2 white, 3 black, 4 native American, 5 others.**

**• education (EDUCYR): years of education.**

**• health (MNHLTH31): 5 excellent, 6 very good, 7 good, 8 fair, 9 poor.**

**• limitation (WLKLIM31): equal to 1 if health limits physical activity.**

**• region (REGION22): 2 northeast, 3 mid-west, 4 south, 5 west.**

**• private (PRIVAT22): equal to 1 if individual has private health insurance.**

**• visits.hosp (OPTOTV22): equal to 1 if at least one visit to hospital outpatient departments.**

**• diabetes (DIABDX\_M18): equal to 1 if diabetic.**

**• stroke (STRKDX): equal to 1 if stroke.**

**• cancer (CANCERDX): equal to 1 if cancer.**

**• income: income (000's).**

**Treat health, race, limitation, and region as a categorical variable when conducting analysis.**

**1. Import the MEPS data. Browse the top 10 rows, and then take a screenshot.**

코드 설명:

2022년 FYC 데이터를 d2022로 저장한 후 위에 언급된 14개의 변수만 선택하여 d2022\_1을 생성하였다. 이후 mutate를 이용해 변수들의 이름을 변경하고 결측값을 처리하는 과정을 거쳤다. 이때 binomial 변수들은 ‘no’응답이 2로 기입되어 있기에 이를 0으로 변경하였고 일부 변수를 범주형 변수로 변환하였다.

결과:

‘김연주\_p4\_분석결과.pdf’ 파일 참고

**2. We want to develop a prediction model for having a visit to hospital outpatient department using other 13 variables. Keep in mind that we want to treat health, race, limitation, and region as a categorical variable.**

**a. First, develop a prediction model using all 13 variables as predictor and interpret the results.**

코드 설명:

Visits\_hosp를 반응변수, 나머지 13개의 변수를 예측변수로 하는 로지스틱 회귀 모델 fit1을 생성하였다. 이후 각 회귀계수의 exp 값을 구하여 각 계수의 오즈비를 계산하였다.

결과:

Intercept는 -3.8728, exp(-3.8728)=0.0208로 이는 모든 독립 변수가 기준 상태일 때 병원 방문 확률을 의미한다. Bmi가 1 증가하면 병원 방문의 확률이 exp(0.0145)=1.1015배가 된다. 같은 방법으로 계산하면 나이가 1세 증가할 때 병원 외래 방문 확률이 약 2.06% 증가하고, 남성은 여성에 비해 병원 외래 방문 확률이 약 33.5% 낮다. 백인의 병원 외래 방문 확률은 기준 인종에 비해 약 1.93배 높고 흑인의 병원 외래 방문 확률은 기준 인종에 비해 약 1.57배 높다. 원주민의 병원 외래 방문 확률은 기준 인종에 비해 약 1.36배 높으며 기타 인종의 병원 외래 방문 확률은 기준 인종에 비해 약 1.90배 높다. 교육 연수가 1년 증가할 때 병원 외래 방문 확률이 약 7.2% 증가하고 건강 상태가 Good인 사람의 병원 외래 방문 확률이 기준 건강 상태(Excellent)에 비해 약 15% 증가하며, 건강 상태가 Fair인 사람의 병원 외래 방문 확률이 약 1.4배 높다. 신체적 제한이 있는 경우 병원 외래 방문 확률이 약 1.6배 높고 중서부 지역에 사는 사람은 기준 지역(북동부)에 비해 병원 방문 확률이 약 1.31배 높으며 남부 지역에 사는 사람의 병원 방문 확률은 기준 지역(북동부)에 비해 40% 낮다. 당뇨병이 있는 사람의 병원 방문 확률은 당뇨병이 없는 사람에 비해 약 28.6% 증가하고 뇌졸중 병력이 있는 사람의 병원 방문 확률은 없는 사람에 비해 약 26.2% 증가하며, 암 병력이 있는 사람의 병원 방문 확률은 없는 사람에 비해 약 1.85배 높다. 마디막으로 소득이 1,000달러 증가할 때 병원 방문 확률이 약 0.1% 증가한다. 이러한 회귀계수들은 health6, health9, region5, private를 제외하고 모두 통계적으로 유의하다. 또한, residual deviance는 11993, AIC는 12037인 것으로 나타났다.

**b. Develop the prediction model using Lasso regression, Ridge regression, and Elastic net regression, respectively and then conduct 20-fold cross validation based on mean squared error criterion. Compare and interpret the results.**

코드 설명:

Glmnet 패키지를 이용하여 lasso, ridge, elastic net regression을 실행하였다. 우선 독립변수와 종속변수를 각각 x, y로 지정하였고 lasso\_cv, ridge\_cv, elastic\_cv를 생성하였다. 이후 최적 lambda와 선택된 변수의 수, 최소 MSE를 도출하였다.

결과:

Lasso regression에서는 MSE가 0.1959, Ridge regression에서는 0.1959, Elastic regression에서는 0.1959가 도출되었다. 분석 결과 세 모델의 MSE가 거의 동일하였다

**c. For each regression, report the value of lambda with the minimum mean crossvalidated error and the number of variables chosen. Report this numerically and visually as we learned.**

코드 설명:

세가지 모델의 최적 lambda, 선택된 변수의 수를 구하였고 이를 시각적으로 나타냈다. MSE 값은 b의 코드를 참고하였다.

결과:

Lasso regression에서는 최적 lambda가 0.000218, 선택된 변수의 수는 21개, 최소 MSE는 0.1959가 도출되었다. Ridge regression에서는 최적 lambda가 0.014679, 선택된 변수의 수는 21개, 최소 MSE는 0.1959가 도출되었다. Elastic regression에서는 최적 lambda가 0.0003972, 선택된 변수의 수는 21개, 최소 MSE는 0.1959가 도출되었다. 시각 자료는 ‘김연주\_p4\_분석결과.pdf’ 참고.

**d. State whether regularization is necessary in this context.**

정규화는 다중공선성 문제가 발생하는 경우와 과적합을 방지하기 위한 의도로 사용된다. 다중공선성이 존재할 경우 변수 간의 상관관계로 인해 회귀계수의 추정이 불안정해질 수 있다. 이러한 경우 ridge 회귀는 계수 축소를 통해 다중공선성을 완화한다. 또한, 변수가 많을 경우 과적합 문제가 발생할 수 있다.

세가지 모델을 비교해보면, 우선 세 모델 모두에서 변수의 선택이 이루어지지 않았기에 과적합 문제가 발생하지 않았음을 알 수 있다. 또한, ridge의 lambda 값이 다른 두 모델의 것보다 크므로 다중공선성 완화를 위한 계수 축소가 상대적으로 강하게 적용되었음을 보여준다. 그러나 애초에 최적 lambda값이 매우 작아 그 차이가 미미하고 MSE도 거의 동일하기 때문에 추가적인 정규화는 필요하지 않다.

**e. Discuss which model is better based on prior analysis.**

세 모델 모두 거의 동일한 분석 결과를 제공하기 때문에 모델 선택에 따른 큰 차이가 발생하지는 않는다. 그럼에도 가장 나은 모델을 고르자면 ridge 회귀 모델이 선택된다. 우선 ridge의 MSE값이 아주 조금 더 작고 그 lambda값이 상대적으로 크기에 규제가 더 강하게 적용되었음을 알 수 있기 때문이다. 또한, lasso와 elastic net의 장점은 변수 선택을 통해 과적합을 방지하고 해석을 용이하게 하는 것인데, 해당 모델에서 변수 선택이 진행되지 않았다. 그렇기에 ridge가 가장 나은 모델로 선택될 수 있다.

**3. We want to develop a classification model for having diabetes. Use BMI, age, gender, education, health care use (as a visit to hospital outpatient department), hypertension, hyperlipidemia, and income as a feature.**

**a. Use logistic regression and then compute the out-of-sample error (misclassification) rate.**

코드 설명:

필요한 변수만 선택한 데이터셋 data를 생성하고 독립변수와 종속변수를 각각 x, y로 지정하였다. 이후 데이터 중 20%를 test data로 정하고 로지스틱 회귀 모델인 logistic \_model을 생성하였다. 이후 misclassification rate을 도출하였다.

결과:

해당 모델의 misclassification rate는 0.1587226이다.

**b. Use KNN and then compute the out-of-sample error (misclassification) rate. Set k = 1, 5, and 50, respectively. Report the results for each k.**

코드 설명:

K=1, 5, 50인 경우에 대해 KNN을 실행하였고 misclassification rate를 계산하였다.

결과:

k=1: 0.2216397

k=5: 0.1720686

k=50: 0.1553861

KNN 모델에서 k값이 클수록 모델의 일반화 성능이 개선된다. K=50일 때의 오분류율이 가장 낮으므로 해당 k값이 가장 적합함을 알 수 있다.

**c. Use classification trees and then compute the out-of-sample error (misclassification) rate.**

코드 설명:

데이터의 20%를 test data로 지정하고 classification tree 모델을 생성하였다.

결과

Misclassification rate: 0.1530029

**d. Compare the results from 3-5 and then describe which model should be chosen.**

코드 설명:

모델별 정확도와 오분류율을 데이터프레임으로 만들어 비교하고 시각화하였다.

결과:

비교 결과 정확도는 classification tree에서 가장 높고 k=1 KNN 모델에서 가장 낮았다. 그에 따라 오분류율은 classification tree에서 가장 낮고 k=1 KNN 모델에서 가장 높았다. 이러한 이유로 classification tree 모델이 선택된다.

**4. We want to create principal components using the following variables (BMI, age, gender, education, health care use [as a visit to hospital outpatient department], hypertension, hyperlipidemia, diabetes, and income).**

**a. First, describe why we need to conduct this analysis.**

주성분 분석이 필요한 이유는 다음과 같다.

1. 다중공선성 문제 해결: 이 데이터에는 변수 간의 높은 상관관계가 존재할 수 있다. 다중공선성은 회귀 분석이나 분류 모델의 성능을 저하시킨다. 주성분 분석은 기존 변수들을 선형 결합하여 상관관계가 없는 새로운 주성분으로 변환함으로써 다중공선성을 완화한다.
2. 차원축소: 주성분 분석은 다수의 상관된 변수를 소수의 주성분으로 변환하여 차원을 줄이는 방법이다. 이렇게 하면 데이터셋의 복잡성을 줄이고 해석을 더 용이하게 만들 수 있다​.
3. 정보 유지: 주성분 분석은 데이터의 변동성을 최대한 유지하면서 새로운 주성분을 생성한다. 첫 번째 주성분은 가장 많은 변동을 설명하고, 두 번째 주성분은 나머지 변동 중 최대치를 설명하는 방식으로 계속된다​
4. 고차원 데이터 해석: 변수 수가 많거나, 관찰 수에 비해 변수가 많은 상황에서는 데이터의 패턴을 찾기가 어려울 수 있다. PCA는 의미 있는 정보를 유지하면서 더 낮은 차원에서 데이터를 분석할 수 있게 돕는다
5. 과적합 방지: 고차원 데이터에서는 모델이 과적합될 위험이 높아진다. 주성분 분석을 통해 차원을 줄이면 노이즈를 줄이고, 모델의 일반화 성능을 향상시킬 수 있다​

**b. As we learned, compute principal components and see how much variations are explained by the first 2 principal components.**

코드 설명:

데이터를 표준화하고 PCA를 실행하였다. 첫번째와 두번째 주성분이 설명하는 분산의 비율을 수치적, 시각적으로 나타내었다.

결과:

첫번째 주성분은 변동의 19.079%, 두번째 주성분은 전체 변동의 16.905%를 설명하여 두 주성분은 35.984%의 변동을 설명한다.

**c. Also, take a look at the first 2 principal components and interpret each principal component.**

PC1: 주요 로딩값 중 visits\_hosp는 -0.360, stroke는 -0.344, cancer는 -0.403, diabetes는 -0.409, age는 -0.548이다. 이는 소득이 양의 방향으로 작게 기여하고 당뇨병, 나이, 병원방문, stroke, cancer가 음의 방향으로 기여함을 보여준다. 즉, 경제적 수준이 높을수록 질병 관련 수치가 낮아지는 것이다.

PC2: 주요 로딩값 중 visits\_hosp는 -0.36, education은 0.583, income은 0.551, cancer는 0.323이다. 이는 교육 수준과 소득이 가장 큰 양의 로딩값을 가지며 나이과 병원 방문이 일부 기여하면서 고소득층의 의료 서비스 이용과 연관된 패턴을 설명한다. 암 관련 수치가 높을수록 병원 방문 횟수와 당뇨병 및 뇌졸증 수치가 낮아지는 패턴이 의아한 결과이다.

**d. Finally, create a figure showing how much each variable contribute to the first 2 principal components**

코드 설명:

PC1, PC2의 변수별 기여도를 데이터프레임으로 나타내고 각각 시각화하였다.

결과:

‘김연주\_p4\_분석결과.pdf’ 참고.