**Take home exam**

**Name: 이지은**

**ID: 222DBG04 빅데이터분석학**

**Prediction Model building and assessing prediction model [add your sas code, if used to answer any of the following questions]**

Q4. Use the demo\_data.xlsx and build the prediction model to detect individuals with diabetes vs. without diabetes

**Development of prediction model**

1. Describe the procedure you used to build your model in sufficient detail, so that a reader could replicate your approach *[hint*: investigators usually build a prediction model including strong risk factors of outcome as independent variables. For identification of risk factors of outcome, investigators either use *a prioi* knowledge including known risk factors of outcome or use data exploration including factors significantly associated with outcome]

저는 일단 리스크 팩터들의 조합을 기존 연구에서 찾았습니다. 아래는 선행 연구 부분입니다.

The prevalence of diabetes increased with age and body mass index and increased inversely with energy expenditure in both males and females. Current and former smokers were associated with a higher prevalence of diabetes. No effect was observed in regular or former drinkers. Prevalence of diabetes increased inversely with income, especially among women. Women who were single and 35 to 64 years old had a higher prevalence of diabetes than women of the same age who were married. The prevalence of diabetes was not found to be related to the level of education. Urban or rural residence was not found to have an effect on the prevalence of diabetes.[Risk factors for diabetes mellitus by age and sex: results of the National Population Health Survey]

There is substantial evidence from a number of cross-sectional and prospective studies that certain ethnic groups, particularly Mexican Americans and African-Americans, have an increased risk of developing of type 2 diabetes when compared with the U.S. population as a whole. Furthermore, there is some evidence that cultural influences may be one factor that increases the risk of disease development. For example, there is an increased incidence of obesity in these populations, which may, in turn, lead to insulin resistance. It appears, however, that genetic factors are also important, as demonstrated by recent analyses that adjusted for factors such as BM1 and found that the risk of developing type 2 diabetes is still elevated in Mexican Americans. Nevertheless, even when the effects of ethnicity and obesity are recognized, the strongest predictors of type 2 diabetes are elevated fasting insulin concentrations and low insulin secretion.[ Epidemiology of Type 2 Diabetes: Risk Factors]

위의 선행 연구들을 참고하여 2형 당뇨병에 대한 분석에 사용할 리스크 팩터들을 선정하였고, 그 조합은 다음과 같습니다: age bmi region sex race

**PROC** **IMPORT** OUT= demo DATAFILE= "C:\Users\USER\Downloads\demo\_data (4).xlsx"

DBMS=xlsx REPLACE;

GETNAMES=YES;

**RUN**;

**proc** **contents**; **run**;

**proc** **logistic** descending;

class race region sex;

model diabetes = age bmi region sex race;

output out= pred p=prob; **run**;

저는 선행연구로 리스크 팩터 조합을 찾았지만 변수선택 코드도 첨부합니다.

**proc** **logistic** data=demo;

model diabetes= sampl strata psu smsa location houssiz age height weight bpsystol bpdiast tcresult tgresult hdresult hgb hct tibc iron hlthstat heartatk sizplace finalwgt leadwt corpuscl trnsfern albumin vitaminc zinc copper porphyrn lead female black orace fhtatk hsizgp hsiz1 hsiz2 hsiz3 hsiz4 hsiz5 region1 region2 region3 region4 smsa1 smsa2 smsa3 rural loglead bmi highbp /selection=forward; /\* 전진 선택법 \*/

**run**;

1. Describe your finale model in sufficient detail and write the equation for the model.

범주형 변수인 race, region, sex를 갖고 age, bmi, female변수를 함께 사용하였습니다.

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명

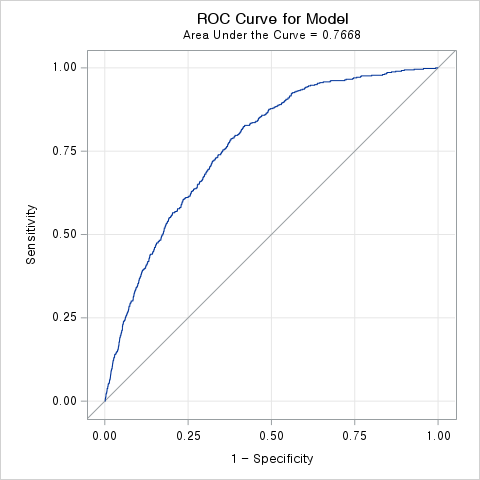
Model의 방정식은 equation=age\*0.0584+bmi\*0.0706+region\_MW\*(-0.019)+ region\_NE\*0.00456+ region\_S\*0.1063+sex\*0.0304+race\_black\*0.2512+race\_other\*0.0768 -7.8242로 도출되었습니다.

**Assess your prediction model**

1. Assess your prediction model; report the discriminatory accuracy; report area under the ROC curve, calibration [Hosmer-Lemneshow statistics] and interpret the results.

텍스트, 스크린샷, 폰트, 라인이(가) 표시된 사진

자동 생성된 설명텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명

**proc** **logistic** descending;

class race region sex;

model diabetes = age bmi region sex race;

output out= pred p=pred\_dia; **run**;

**proc** **corr** data=pred;

var diabetes pred\_dia; **run**;

피어슨 상관계수는 0.20952로 높지 않게 나왔으며, 그 아래 그래프에서 실제로 모델을 통해 예측된 결과와 실제값을 볼 수 있었습니다. roc커브를 통해 모델의 성능을 확인할 수 있었으며, auc=0.7668의 준수한 성능을 보여주는 것을 알 수 있습니다.

**proc** **logistic** descending data=demo;

class race region sex;

model diabetes = age bmi region sex race/ outroc=ROCData; **run**;

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명

**proc** **logistic** descending;

class race region sex;

model diabetes = age bmi region sex race/lackfit;

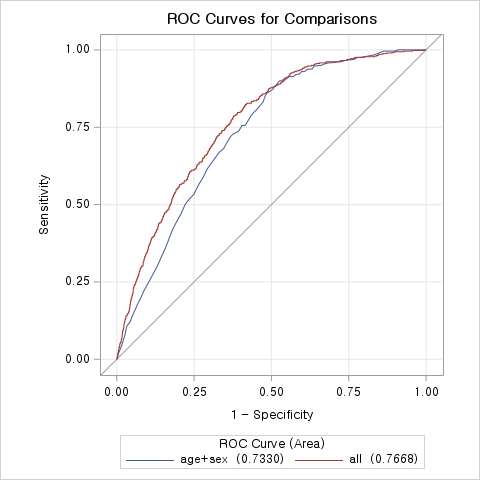
**run**;

Hosmer-Lemneshow statistics을 통해 10개의 그룹에서 실제값과 유사하게 예측이 잘 일어나고 있는 것을 확인할 수 있습니다.

**Comparison of prediction models**

Assumes that you want to see if your prediction model developed in Q4.a-b is better prediction model or not, compared to the reference model with age and sex.

1. Create the table comparing ROC curves across two models and net reclassification and interpret the results (assume low risk :<10% predicted probability, medium risk :10- <20% predicted probability, ≥20% predicted probability; refer to the table in the "Calibration\_netclassification\_validation.pptx” slide 40 page “)



**proc** **logistic** descending data=demo plots(only)=(roc);

class race region sex;

model diabetes = age bmi region sex race/nofit;

roc "age+sex" age sex;

roc "all" age bmi region sex race;

roccontrast reference("age+sex") / estimate; **run**;

roc곡선을 통해 선택된 리스크 팩터를 모두 사용하는 것이 age와 sex변수만을 사용한 모델보다 성능이 개선되었음을 알 수 있었습니다. 또한 auc도 모두 사용한 모델이 0.7668로 더 큰 값을 가집니다.

|  |  |  |
| --- | --- | --- |
|  | p-value | AUC |
| All(age+bmi+region+sex+race) | <.0001 | 0.7688 |
| Age+sex | <.0001 | 0.7330 |

|  |  |  |
| --- | --- | --- |
| All(age+bmi+region+sex+race) | **모델 1 예측 negative** | **모델 1 예측 positive** |
| True Negative |  |  |
| True Positive |  |  |

|  |  |  |
| --- | --- | --- |
| Age+sex | **모델 2 예측 negative** | **모델 2 예측 positive** |
| True Negative |  |  |
| True Positive |  |  |

1. Is your prediction model better than the reference model? what is your conclusion ? justify your answer with the results you obtained from Q4-d

Auc로 판단하였을 때 제가 제안하는 모델이 0.7688로 기존 ref 모델보다 더 좋은 성능을 보여주고 있습니다.

1. What would be the next step? In other words, what is the main limitation of assessing the prediction model accuracy using the data used to develop the prediction model?

train데이터를 모델 평가에 그대로 사용하는 것은 과적합 문제가 발생할 수 있어서 이를 방지하기 위해서는 데이터를 training 데이터와 test 데이터로 분리한 후 test 데이터를 사용해서 모델의 성능을 평가해야 합니다.