<선형회귀분석, 로지스틱 회귀분석 보고서>

정보융합학부 2018204002 박정재

1. Odds 사례

Odds는 간단히 말해 승산으로서, 어떤 상황에 대해 성공확률이 p라고 하면 성공대비 실패의 확률비율을 말한다. 이에 관한 예시로 세브란스·대구동산병원 연구팀의 연구결과를 가져왔다. 당뇨병 환자가 중증 저혈당을 앓고 있는 경우 그렇지 않은 환자에 비해 사망위험도가 4.3배, 치매를 앓고 있는 사람의 경우 2배, 두 질병 모두 앓고 있는 사람의 경우 5.1배 상승했다고 한다. 그렇다면 이 결과들을 Odds의 개념으로 보면 순서대로 각각의 세가지 상황들은 질병에 걸렸을 때 사망할 확률이 걸리지 않았을 때의 사망확률보다 각각 4.3, 2, 5.1배 높다는 것을 의미한다. 따라서, 당뇨병 환자의 경우 해당 두개의 질병에 걸렸을 때 사망위험이 매우 높아진다는 것을 알 수 있다.

[ 출처-<https://sports.chosun.com/news/ntype.htm?id=202204120100085080005515&servicedate=20220411>]

2. 선형회귀분석, 로지스틱 회귀분석 분석 결과

2.1) 선형회귀분석

1>데이터 수집 및 전처리

텍스트이(가) 표시된 사진

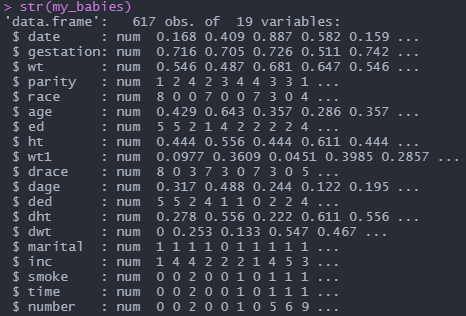
자동 생성된 설명

텍스트, 창문이(가) 표시된 사진

자동 생성된 설명텍스트이(가) 표시된 사진

자동 생성된 설명

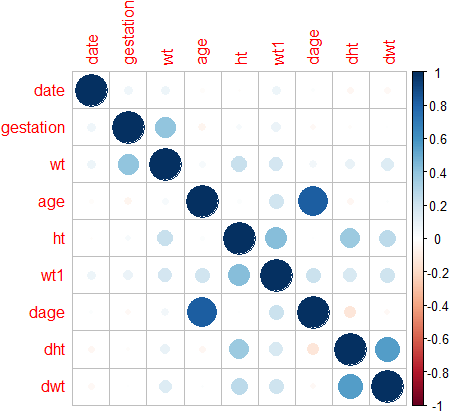
텍스트, 명판, 스크린샷이(가) 표시된 사진

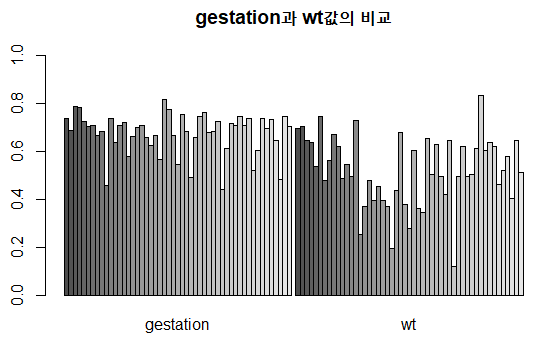
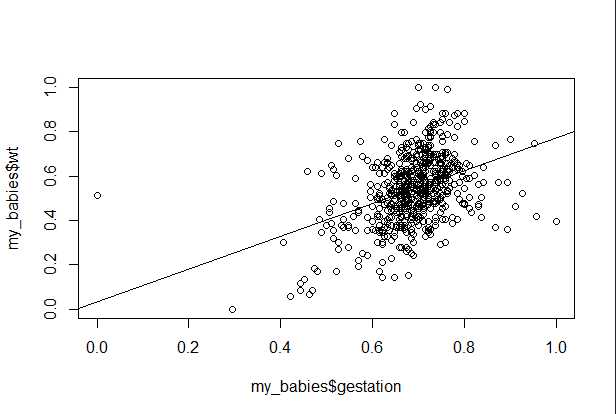
자동 생성된 설명

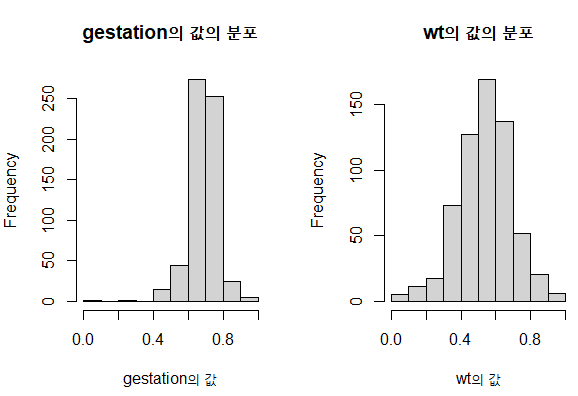
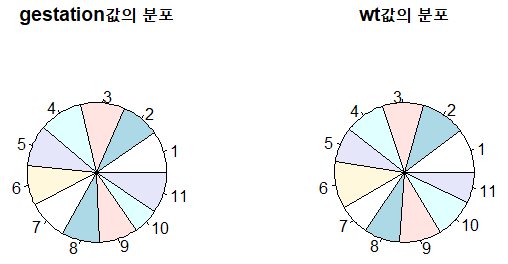
선형회귀분석에 사용한 babies 데이터셋은 23개의 attribute와 1236개의 tuple로 구성되어 있는 데이터이며, 신생아와 부모의 정보들을 담고 있다. 우선 의미 없는 변수들(id, pluralty, outcome, sex)을 제거하고, 남아있는 변수들의 값들 중 의미 없는 값들을 제거했다. 그 후, 범주형 변수들을 수치형으로 변환 후 구간을 다시 나누어 주었고 수치형 데이터들의 값들을 범위가 같아지게 정규화 해줬다.

2> 탐색적 데이터 분석

텍스트이(가) 표시된 사진

자동 생성된 설명





먼저, 수치형 변수들 사이의 상관관계를 알아보기 위해 corrplot을 사용하였고, 변수 gestation과 wt간의 상관관계가 높다는 것을 알아냈다. 그 후, scatter plot을 통해 두변수가 높은 선형적 관계에 있다는 것을 알아냈고, barplot, piechart, histogram 차트들은 두 변수사이의 선형적 관계를 더 뒷받침해준다. 이 사실들을 통해, 종속변수를 wt(아기의 몸무게)로 정해주었다.

3>학습모델 구축

텍스트이(가) 표시된 사진

자동 생성된 설명

wt(아기의 몸무게) 예측을 위한 선형모델 구축하기

3.1> 단순선형회귀 모형 해석

텍스트이(가) 표시된 사진

자동 생성된 설명

우선, p-value 값으로 봤을 때 회귀계수 모두 유의하다는 것을 알 수 있다. 따라서, 본 회귀모델을 통해 아이의 몸무게는 임신기간에 따라 0.73배 높아질 수 있음을 알 수 있다. 또한, 이 선형모델은 R-squared 값에 의해 15.89% 정도의 설명력이 있고, Adjusted R-squared값에 의해 15.75%의 설명력이 있다.

3.2> 다중선형회귀 모형 해석

텍스트이(가) 표시된 사진

자동 생성된 설명텍스트이(가) 표시된 사진

자동 생성된 설명텍스트이(가) 표시된 사진

자동 생성된 설명텍스트이(가) 표시된 사진

자동 생성된 설명

결국, 유의수준(5%)를 기준으로 gestation, parity, race, ht, dwt, smoke, number 모두 유의미한 변수로 설명이 된다. 또한, R-squared 값에 의해 26.94% 정도의 설명력이 있고, Adjusted R-squared값에 의해 26.1%의 설명력이 있다.

3.3> 모델비교 및 모형진단

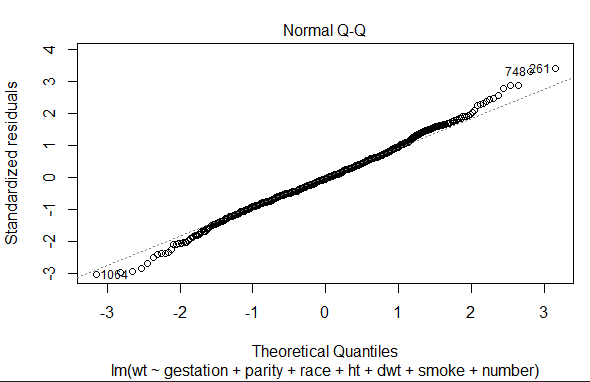
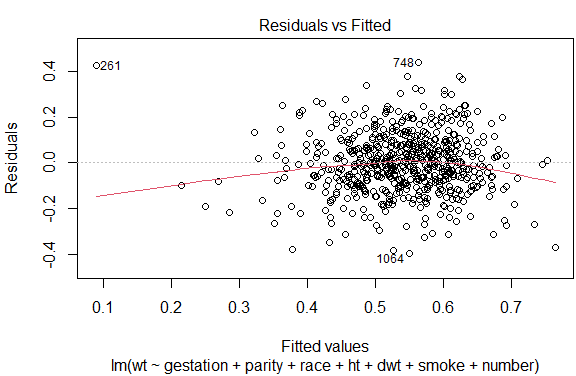
<모델비교>

 텍스트이(가) 표시된 사진

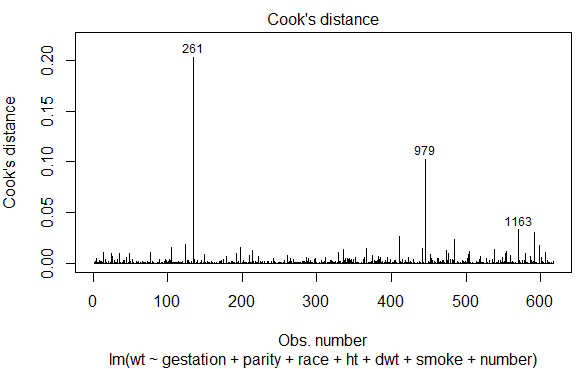
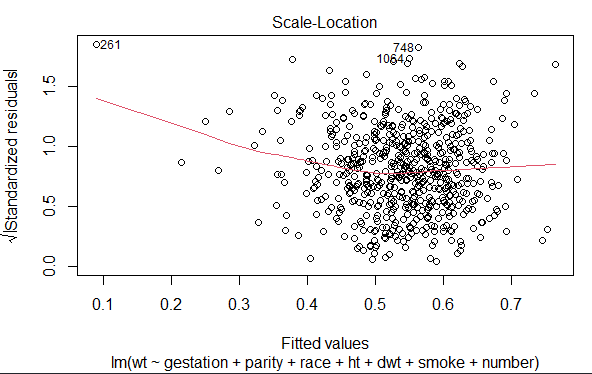
자동 생성된 설명

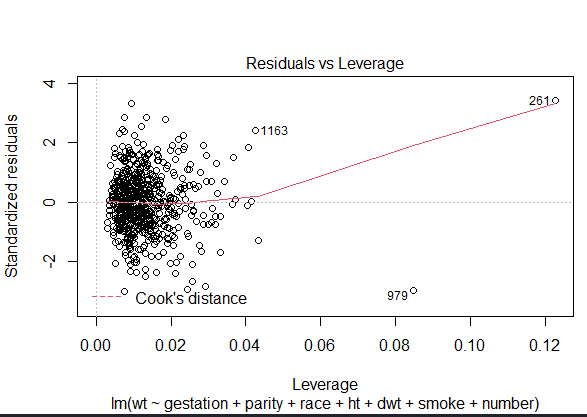
단순선형회귀 모델 model 과 다중선형회귀인 multi\_model 간 분산의 차이가 있는지에 대한 가설검정 결과 p-value가 2.2e-16으로 0에 가깝기 때문에, multi\_model 선형모델이 유의하다는 것을 알 수 있다.

<모형진단>



우선 왼쪽 그래프의 실제(y)값과 예측 값 사이의 잔차를 확인해 보았을 때, 261, 748번 쨰 값들은 예측한 값보다 훨씬 좋았고 1064번째 값은 예측값보다 훨씬 안 좋았음을 알 수 있다. 다음으로 QQ-plot을 봤을때, 대부분 정규성에 잘 맞았지만 261, 748번째값들은 정규성에 잘 맞지 않는것으로 해석 되었다.





Scale-Location을 봤을 때, Y값이 커짐에 따라 잔차가 커지며 이상치들이 발견되고 있다고 해석 할 수 있다. 다음으로 Cook’s distance를 보면 261, 979, 1163 번째 값들이 회귀직선의 기울기에 꽤 큰 영향을 준다는 것을 알 수 있다. 마지막 그래프를 보면 회귀직선에 큰 영향을 주던 값들이 매우 극단적인 값을 가졌음을 알 수있다. 따라서, 이 데이터들은 걸러내는 것이 유의미 하다는것을 나타낸다.

4> 결과 해석

텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

자동 생성된 설명

RMSE는 MSE의 제곱근의 값으로, RMSE를 제곱하게되면 MSE의 값을 알 수 있고, MAE의 값도 명시된 값을 통해 알 수 있다. 이 때, multi\_model이 model보다 예측오차값들이 작다는 사실로보아 multi\_model이 model보다 예측력이 좋다는 것을 알 수 있다.

2.2) 로지스틱회귀 모형 분석

1> 이진분류 성능평가 함수 정의

텍스트이(가) 표시된 사진

자동 생성된 설명

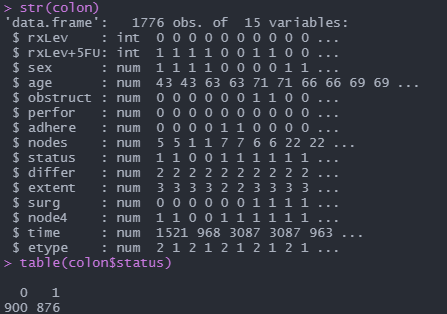
로지스틱회귀 모형의 결과 분석을 해줄 이진분류 함수를 정의하여 만든다.

2> 데이터 로드 및 전처리

텍스트이(가) 표시된 사진

자동 생성된 설명 텍스트이(가) 표시된 사진

자동 생성된 설명



로지스틱 회귀분석에 사용한 colon데이터셋은 16개의 attribute와 1858개의 tuple로 구성되어 있는 데이터이며, 대장암 환자들의 정보들을 담고 있다. 우선 결측 값들을 제거해 주었고, 범주형 변수를 더미변수로 전환해 주었다. 그 후, 의미 없는 변수들(id, study, rx-Obs)를 제거하고 종속변수로 status를 선택하였다. 이 때, status가 0인경우가 900명 1(대장암 재발 또는 사망)인 경우가 876임을 명시해 놓았다.

3> 데이터 학습/테스트 구분

텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

자동 생성된 설명

데이터를 학습 데이터와 테스트데 이터로 구분하기위해 test\_id를 만들었고, 테스트를 전체횟수의 96%만큼 할당했다. 그 후 나누어진 데이터들을 보면 학습데이터가 71, 테스트데이터가 1705개임을 알 수 있다.

4> 로지스틱회귀 모형 구축 및 예측 수행

텍스트이(가) 표시된 사진

자동 생성된 설명

우선 모든 독립변수들로 status에 대해 예측할 모델을 학습 데이터를 이용하여 만들고, 테스트 데이터를 이용하여 예측된 확률이 0.5를 기준으로 0혹은 1의 값을 갖도록 pred\_class를 만들어준다. 그 후, 실제 값과 비교하여 값이 같은 경우와 다른 경우들을 이진 분류함수로 분류하여 예측을 수행한다.

텍스트이(가) 표시된 사진

자동 생성된 설명

예측 정확도는 86%정도가 나왔고, F1값은 0.844정도가 나왔다.

5> 변수 선택법 적용

텍스트이(가) 표시된 사진

자동 생성된 설명

변수 선택법 forward selection과 step selection을 사용하여 정확도를 향상시켜보았다. 변수의 개수가 적어 두방법의 결과값은 같았다.

6> 결과 해석







텍스트, 장치, 게이지이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

자동 생성된 설명

우선 모든 변수선택법 적용전과 후의 confusion matrix는 다음과 같고, 예측오류값들이 줄어들었음을 한눈에 알 수 있다. 이 때, 이진 분류 함수를 통해 구해진 값들에서 recall의 값은 TPR, 예측정확도 값은 Precision, F1은 F1값이다. 예측 정확도와 recall, F1값이 상승함을 통해 변수 선택법으로 유의미 하지 않은 변수들을 제거함으로써 예측력이 높아졌음을 알 수 있다.