

## 데이K0076-01 일반화선형모형 시험문제

2020년 10월 21일

과제는 반드시 R프로그램을 이용하여 프로그램과 결과를 제출하기 바랍니다(그렇지 않으면 점수가 없습니다).

1. 다음은 호흡기질환 환자들에게 시험(test)약과 위약(placebo)을 비교 처치한 결과이다. R프로그램을 이용하여 다음의 질문에 답하시오(30점)

처리	개선됨(favor)	개선 안됨(unfavor)	계
위약	16	48	64
시험약	40	20	60
계	56	68	124

- 1) 위약과 시험약의 개선 비율의 차와 비율에 차에 대한 신뢰구간을 구하고 연관성을 해석하시오.
  - 2) 위약과 시험약의 개선에 대한 상대위험도(relative risk)와 상대위험도에 대한 신뢰구간을 구하고 연관성을 해석하시오.
  - 3) 오즈비를 구하고 해석하시오
  - 4) 오즈비에 대한 신뢰구간을 구하고 연관성을 해석하시오.
  - 5) 위의 각 셀에 대한 기대도수의 추정치를 구하시오.
  - 6) 호흡기질환 환자들에 대한 처리와 반응이 관련이 있는지에 대한 귀무가설을 적고 그에 대한 피어슨통계량과 가능도비통계량을 찾고 연관성에 대해 가설검정을 하시오.
2. 포아송회귀분석과 음이항 회귀분석을 이용하여 주어진 데이터(birthdata.xlsx) 141명의 여성의 나이를 통해 자녀의 수를 예측하고자 한다. R프로그램을 이용하여 다음 문항에 답하시오. (30)
- 1) 위의 자료를 이용하여 포아송회귀모형과 음이항회귀모형식을 제시하시오. 어느 경우에 포아송회귀모형과 음이항회귀모형식을 사용하는지 설명하고 이 데이터의 구조를 살펴보고 어떠한 모형이 더 나은지 선택하고 그 이유를 설명하시오. (5)
  - 2) 위의 적합된 두 모형을 활용하여 각각에 대해 나이가 40세인 여성의 평균 자녀 수 예측값을 구하시오. (5)
  - 3) 포아송 회귀모형을 이용하여 자녀 수에 대한 나이의 효과를 설명하고 그 효과에 대한 신뢰구간을 구하시오. (5)
  - 4) 적합된 음이항 회귀모형을 이용하여 여성의 평균 자녀의 수는 나이와 독립인지를 검정하시오. (5)
  - 5) 엄마의 나이에 따른 자녀의 수에 대한 그래프를 그리시오. (5)
    - 산점도를 그리시오
    - 포아송회귀 모형에 대한 예측값(빨간색)과 음이항 회귀모형에 대한 예측값(파란색)을 실선과 점선으로 표시하여 산점도 위에 연결하여 표현하시오.
  - 6) 가능도비(LRT) 검정법이 사용가능한 경우가 언제인지 설명하고 가능도비 검정법을 사용하여 포아송회귀모형과 음이항회귀모형식 중 어느 모형이 나은지 선택하고 설명하시오.
3. 주어진 데이터 diabetes.xlsx는 호주 New South Wales주의 지역들에서 당뇨병(diabetes)에 기인한 사망자의 수를 다룬다. 엑셀에 포함된 변수들은 다음과 같다. (40)

Variables:      코드 또는 내용

gender            Male, Female

age	<25, 25-34, ..., 85+
deaths	사망자의 수
popn	인구의 크기
agemidpt	연령의 중간값

- 1) 인구의 크기가 주어질 때 성별과 연령에 따른 당뇨병 환자의 사망자 수를 예측하는 포아송회귀모형을 제시 하시오. 다음과 같이 연령을 연속변수로 사용하는 것이 나은지 범주형으로 사용하는 것이 나은지 선택하여 최종 모형을 선택하시오. (3개의 모형 제시 바람) (15점)
  - 성별은 범주, 연령(연령의 중간 값 이용)은 연속변수로 1차항만 사용
  - 성별은 범주, 연령(연령의 중간 값 이용)은 연속변수로 1차항과 2차항 사용하는 경우
  - 성별은 범주, 연령을 범주형으로 사용하는 경우(기준수준으로 45-54세를 이용하기 바람)
 또한 연령과 성별 중 어느 변수가 당뇨병 환자의 사망자 수에 영향을 주는 지 근거를 제시하여 설명하시 오.(5)
- 2) 연령 대비 남녀 별 10만 명 당 사망자 수를 나타내는 산점도 그래프(남자는 파란색, 여자는 빨간색)를 동시 에 나타내고 성별 각각에 대한 예측값을 선으로 연결하시오. (5)
- 3) 남성이고, 연령은 55-64세, 인구수는 500000명 일 때, 여성이고, 연령은 75-84세, 인구수는 70000명 인 경우 각각 당뇨병환자의 수의 예측값과 신뢰구간을 구하시오. (5)
- 4) 당뇨병 환자의 사망률을 나타내는 모형을 제시하시오. 이 모형은 1)에서 선택된 모형과 차이는 무엇인지 설명 하시오.(5). 이를 이용하여 남성이고, 연령은 55-64세, 인구수는 500000명 일 때, 여성이고, 연령은 75-84세, 인구수는 70000명 인 경우 각각 당뇨병환자의 사망률에 대한 예측값과 신뢰구간을 구하시오. (5)