Homework

`Wage`: Mid-Atlantic 지역 내 3,000명의 남자 근로자 임금 자료[1]

- 데이터 설명
  - 설명 변수 $V_1, \ldots V_9$: 조사한 연도, 근로자의 나이, 결혼 여부, 인종, 교육 수준, 지역, 직업 소분류 (산업 / 정보), 건강 상태, 보험 유무
  - 반응 변수 V_{10}, V_{11}: log(wage), wage
- 실습
  주어진 `Wage` 자료에 대한 탐색적 자료 분석 및 통계 검정 진행.

  1. 기초통계와 탐색적 자료분석
     - 자료에 결측치 여부 확인.
     - 결측치가 존재할 경우, 행을 지우고 아래 분석을 진행.
     - 연속형 설명변수에 대해, 요약통계량 (평균, 분산, 중앙값, 최소값, 최대값) 및 boxplot 살펴보기.
     - 범주형 설명변수에 대해, 해당 범주에 속하는 비율과 해당 범주에 따른 wage의 분포 살펴보기.
     - 반응 변수 중 logwage은 wage 변수에 로그를 취한 값이다. 이 둘의 히스토그램을 각각 그려서 비교하여라.
     - 연속형 설명변수와 logwage에 대한 pairwise plot을 그려보고 각 변수마다, 그리고 설명변수와 반응변수 내에 선형적인 관계가 있는지 그림을 통해 설명하여라. 마찬가지로, 연속형 설명변수와 wage에 대해 pairwise plot을 그려보고 비교하여라.
     - 범주형 설명변수인 직업 소분류 변수에 대해 자료를 그룹화하고, 각 그룹마다 pairwise plot을 통해 차이가 있는지 살펴보아라. 다른 범주형 변수인 교육 수준에 대해 마찬가지로 진행하여 살펴보아라.
     - log(wage)의 중앙값을 계산하고, 중앙값보다 크면 $y_i = 1$, 작거나 같을 경우 $y_i = 0$으로 자료에 대한 라벨링을 진행하여라. 그리고 y값을 그룹을 의미하는 변수로 사용하여 연속형 설명변수에 대한 pairwise plot과 범주형 변수에 대한 기초통계를 진행하여라.
     - 근로자의 결혼 상태외 건강 상태에 대해 2차원 분할표를 구하고, 의미를 해석하여라.
     - 교육수준과 직업 소분류에 대한 2차원 분할표를 구하고, 의미를 해석하여라.

[1] `http://thedataweb.rm.census.gov/TheDataWeb` 참고.

2. 통계 검정

- 미국 지역 내 logwage의 평균은 4.7으로 알려져있다. 이 때 Mid-Atlantic 지역 내 근로자는 전체 미국 logwage 평균 임금보다 더 많이 임금을 받는다고 판단할 수 있는가? logwage가 정규분포라는 전제 하에서 적절한 가설을 세우고, 유의수준 5%의 검정을 진행해보아라.
- 위와 같은 검정을 wage 변수에 대해서도 진행하여라. 이 때 미국지역 내 wage의 평균은 $\exp(4.7)$이다. 두 검정을 진행하고 결과에 대해 논의를 진행하여라.
- 직업 소분류에 따른 logwage의 모평균이 차이가 있는지 비교하고자 한다. 적절한 가설과 검정통계량을 사용하여, 유의수준 5%의 검정을 진행하고, 결과에 대해 논의를 진행하여라.
- 건강 상태에 따른 llogwage의 모평균이 차이가 있는지 비교하고자 한다. 적절한 가설과 검정통계량을 사용하여, 유의수준 5%의 검정을 진행하고, 결과에 대해 논의를 진행하여라.
- 보험 유무에 따른 logwage의 모분산이 차이가 있는지 비교하고자 한다. 적절한 가설과 검정통계량을 사용하여, 유의수준 5%의 검정을 진행하고, 결과에 대해 논의를 진행하여라.