Mock Test for Midterm

Data Science 2020 spring

2020-04-17

7:00 ~ 9:00 PM

Perform the following tasks and **put your answer in a document file (MS word, HWP, or PDF, or any format)**. Submit your **R script** that you use to find the answer with **the document**. You can choose to answer in your proficient language from English and Korean. All the answers should be based on findings from the given dataset.

(Korean Translation) 아래의 질문에 대한 답을 문서파일로 작성(MS word, 한글파일, PDF 또는 원하는 형식 가능)하고 답을 찾기 위해 사용한 R script 파일과 함께 edmodo에 제출하세요. 답은 한글과 영어 중 편한 언어로 작성하면 됩니다. 모든 답은 데이터로부터 발견한 사실에 기반해야 합니다.

(주의) 9시까지가 시험이고 이 시간을 지나면 (9시 정각이 되면) 답안을 제출하지 못하게 됩니다. 답안이 제출되지 않은 경우 자동 0점 처리되고, 어떤 예외도 허용되지 않기 때문에 제시간에 제출하는 것을 주의하세요. 한번 제출한 이후에 시간 내라면 여러 번 제출하는 것은 가능합니다.

You must submit your work to Edmodo in time (~9:00PM). The submission page will be locked after 9PM. Only the work that are submitted to Edmodo successfully in time will be graded. You can submit your work several times during the exam (7PM ~ 9PM) if necessary.

1. Data Preparation and Exploration

For the first set of questions, you will be working with a dataset contained in "insurance.csv" file. The dataset contains information about customers of Health Insurance Company and amount of money they charged to the insurance plan for a year as medical expenses. The data contains 7 variables of "age", "sex", "bmi", "children", "smoker", "region", and "charges". Read the follows for more detailed description of variables.

(Korean Translation) "insurance.csv" 파일에 저장된 데이터를 사용하게 됩니다. 제공되는 데이터에는 미국 건강 보험 회사의 고객 정보가 포함되어 있습니다. 데이터에는 7개의 변수가 있으며 각각 고객의 나이, 성별, bmi, 자녀수(보험으로 보장받는), 흡연여부, 거주지역, 그리고 1년간 총 의료보험 지급액(또는 보험 수령액)을 의미하는 변수 age", "sex", "bmi", "children", "smoker", "region", and "charges"로 되어있습니다. 변수에 대한 자세한 설명은 아래에.

- `age`: This is an integer indicating the age of the primary beneficiary (excluding those above 64 years, since they are generally covered by the government).
- `sex`: This is the policy holder's gender, either `male` or `female`.
- `bmi`: This is the **body mass index (BMI)**, which provides a sense of how over or under-weight a person is relative to their height. BMI is equal to weight (in kilograms) divided by height (in meters) squared. An ideal BMI is within the range of 18.5 to 24.9.
- `children`: This is an integer indicating the number of children / dependents covered by the insurance plan.
- `smoker`: This is `yes` or `no` depending on whether the insured regularly smokes tobacco.
- `region`: This is the beneficiary's place of residence in the U.S., divided into four geographic regions: `northeast`, `southeast`, `southwest`, or `northwest`.

- **Charges:** The customer's total medical expenses charged to the insurance plan for a year

$$\mathrm{BMI} = \frac{\mathrm{mass_{kg}}}{\mathrm{height_m^2}}$$

1. Load the dataset into R as data.frame. Explain how you loaded the data and why you did so.

(Korean Translation) 파일에 저장된 데이터를 data.frame으로 R에 읽어오세요. 어떤 방법으로 읽었으며 왜 그렇게 하였는지 설명하세요.

2. Perform data exploration process on the dataset. Describe what you find during the process and how in the document.

(Korean Translation) 읽어온 data.frame에 대해서 데이터 탐색을 수행하세요. 탐색 과정 중에 파악한 정보가 있으면 설명하고 어떻게 그것을 찾았는지 설명하세요.

3. Normal range of BMI is known to be within range (18.5, 24.9). Group the customer into 3 groups "light", "normal", "heavy" with BMI range (0,18.5], (18.5, 24.9], (24.9, ~), respectively. To which group of customers the customer pays the most and the least? Compare average charges of the groups.

(Korean Translation) BMI(체질량지수)는 18.5에서 24.9사이가 보통이라고 합니다. 고객을 bmi 값 (0,18.5], (18.5, 24.9], (24.9, ~) 에 따라 "light", "normal", "heavy" 세 그룹으로 나누어서 각 그룹에 대한 평균 보험 수령액을 비교하세요. 어떤 그룹이 가장 보험금이 많이 들었고 어떤 그룹이 적게 들었나요?

4. Compare average bmi of male and female customer group. Do you find that bmi distribution are different for male and female group in terms of sample mean and standard deviation? You may draw some plots such as histogram and probability density plot if needed.

(Korean Translation) 남성과 여성 고객의 평균 bmi를 비교해보세요. 두 집단의 평균과 표준편차를 고려할 때, 남성과 여성의 bmi 수치는 다른 분포를 갖는다고 할 수 있나요? 필요하다면 히스토그램이나 확률밀도 함수와 같은 그래프를 그려서 확인해볼 수 있습니다.

5. Seeing "charges" variable, explain how "charges" variable is distributed (e.g. normal distribution, uniformly distributed, skewed, etc.). Describe the distribution of why it is so.

(Korean Translation) 보험수령액 "charges" 변수의 값은 어떻게 분포하고 있나요? 분포에 대해서 설명하고 왜 그런 분포가 나타나는지 설명해보세요.

6. From the dataset, can you say "smoker" customers should pay to the insurance company more than "non-smoker" customer? Explain why you answer so.

(Korean Translation) 데이터를 보았을 때, 흡연 고객이 비흡연 고객보다 보험료를 더 많이 내야 한다고 생각하나요? 이유를 설명해보세요.

7. According to the dataset, can you say more children a customer has, more medical expenses the customer charges to the insurance? Explain your answer using what you have found from the dataset.

(Korean Translation) 데이터를 보았을 때, 더 많은 자녀가 있다면 더 많은 의료비용이 발생한다고 볼 수 있나요? 답을 쓰고 답에 대해서 설명해보세요.

8. Find the customer group A of oldest 10% and group B of youngest 10%, how much more expenses the group A charges to the insurance company compared to expenses of group

B on average?

(Korean Translation) 고객들 중 나이가 가장 많은 10%의 A그룹와 나이가 가장 어린 10% 그룹 B로 나누었을 때, 그룹 A는 그룹 B에 비해서 얼마나 많은 보험 지급액이 많이 발생하나요?

9. Compare average “charges” of male and female customer groups. How are they different?

(Korean Translation) 성별에 따른 평균 보험지급액을 비교해보세요? 어떻게 다른가요?

10. Compare the portion of “smokers” of male and female customer group. Can we say male and female customer have different average medical expense, because they have different portion of “smokers”. Or would you say the difference comes from the gender no matter whether they smokes or not? Give findings that support your answer from dataset and explain your answer with it.

(Korean Translation) 남성과 여성을 비교했을 때 흡연 고객의 비율은 어디가 더 많은가요? 만약 남성과 여성의 보험지급액이 다르다면 그것은 두 성별간의 흡연 고객의 비율이 다르기 때문이라고 설명할 수 있을까요? 그렇지 않다면 흡연 여부와 상관없이 성별의 차이 때문에 발생한다고 할 수 있을까요? 이것에 답하기 위한 정보를 데이터로부터 찾아보고 답을 해보세요.

2. Tidy data

"**automobile.tsv**" contains a data.frame that has information about 396 car models. Perform following tasks or answer question.

(Korean Translation) "**automobile.tsv**" 은 396개의 자동차에 대한 여러 정보를 포함하고 있는 dataset이다. 다음 질문에 답하거나 task를 수행하라.

1. The dataset contains information about 396 sample car models (assuming one observation corresponds to one car model). The attributes (variables) includes "car name", "release year", "fuel efficiency", "weight", and so on. Do you find whether the dataset is tidy or not tidy? Explain your answer with proper reason.

   데이터에 표현된 observation은 각각 하나의 자동차 종류를 나타낸다고 한다. 또한 변수는 자동차의 여러 특징(자동차 이름, 연식, 연비, 무게 등)을 표현한다고 하자. 그렇다면 이 데이터는 tidy 데이터인가 그렇지 않은가? 이유를 설명하시오.

2. If you think the dataset is not tidy, then transform the dataset into tidy one. Save the tidy form of the data into "**automobile.RData**" file with save command. The file should include only one data frame.

   1번의 답이 tidy가 아니라고 한다면, tidy 형태로 변환하시오. 결과는 RData에 저장(save 명령 사용)하여 제출하시오. 파일에는 하나의 데이터 프레임만 저장되어 있어야합니다.