Flu Shot Prediction

신종플루와 계절독감 백신 접종 여부 예측

범주형자료분석팀

이지연 | 심예진 | 조장희 | 조혜현 | 진효주

CONTENTS

1

주제 선정

2

DATA

3

EDA

4

4주차 예고

- **Flu Shot Prediction**

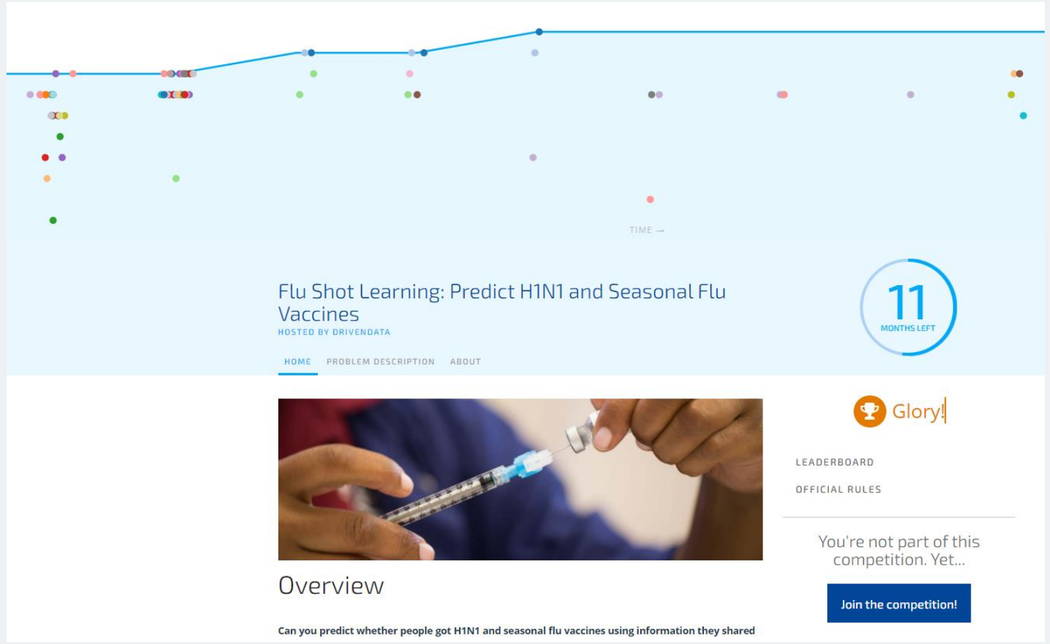
데이터 사이언스를 통해 세계의 사회적 문제들을 다루는 온라인 과제를 주최하는 Driven Data의 분석 과제 중 National 2009년 H1N1 Flu Survey 데이터셋을 통해 신종플루와 계절독감 백신 접종 여부를 예측하는 문제

35 Categorical Features &
2 Labels

설문조사 응답 데이터로

Feature변수와 Label 변수 모두 **범주형 변수!**

범주가 범주했다 2탄…



This is competition……

▪ Multi-Label Classification Problem

H1N1 백신 접종 여부와 계절독감 백신 접종 여부를 모두 예측해야 하는 상황!

Multi-Label Classification

일반적인 분류 모델과 달리 여러 라벨 값을 예측해야 하는 분류

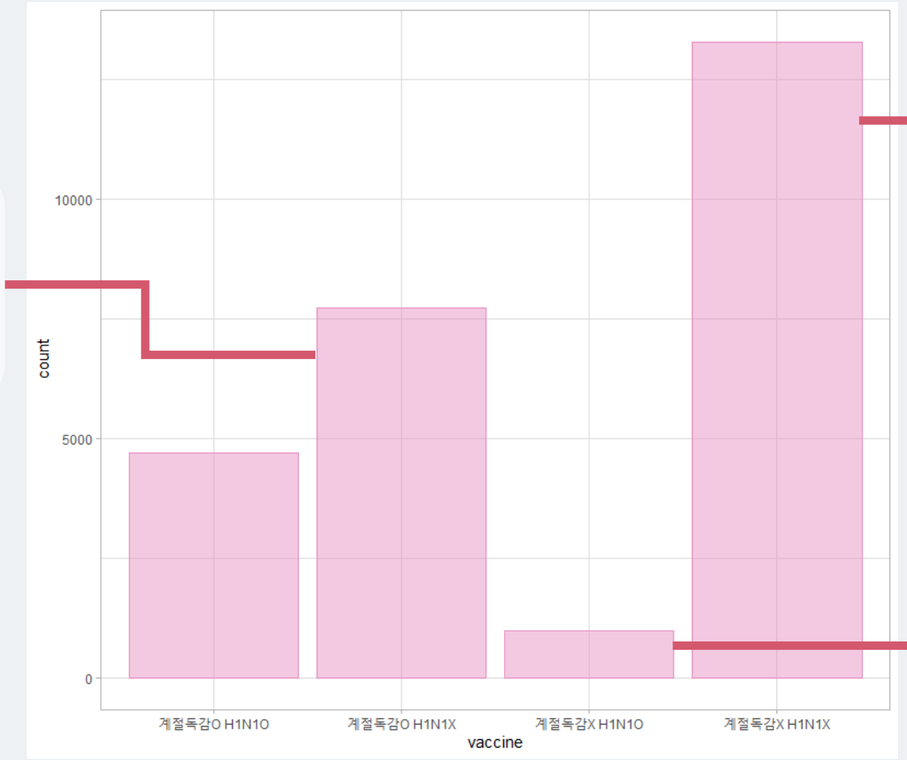
Multi-class classification과 다른 점?

멀티 클래스 분류에서는 각 클래스가 서로 상호 배타적(mutually exclusive)한 관계라면,
멀티 레이블 분류에서는 각 레이블이 다른 분류 문제이지만 서로 관련이 있는 경우

▪ Label 변수

예측해야 하는 종속변수는 계절독감 백신 접종 여부와 H1N1 백신 접종 여부 2가지!

〈각 백신 접종 여부 분포〉



계절독감 백신만 접종한 응답자는 다수

두 백신 모두 접종하지 않은 응답자 가장 많음!

계절독감 백신은 맞지 않았으나 신종플루 백신만 맞은 사람이 가장 적음

▪ 우리의 설문조사 무응답 데이터는 어느 종류에 속할까?

왜 MNAR인가?

1. 의도적인 무응답 가능성 높음

- 선택적 질문이 아니라 응답자들 전원이 모든 질문을 받고 응답한 상황
- 경제적 상황 관련 질문과 같은 응답하기 민감한 질문에 무응답 비율 높음

2. 변수별 결측 여부의 상관성 존재

- 특정 변수에 무응답한 경우 다른 변수에도 무응답한 경우 존재

BUT MNAR은 MCAR 가정과 다르게 대체 방법을 사용하기 쉽지 않음!

합리적 접근 방법 가능한지 시각화를 통해 확인해보자

** **합리적 접근법?** 변수들간 관계를 이용해서 결측치를 채우는 방법 - 도메인 지식 필요 !*

▪ 합리적 접근법을 활용한 결측치 대체

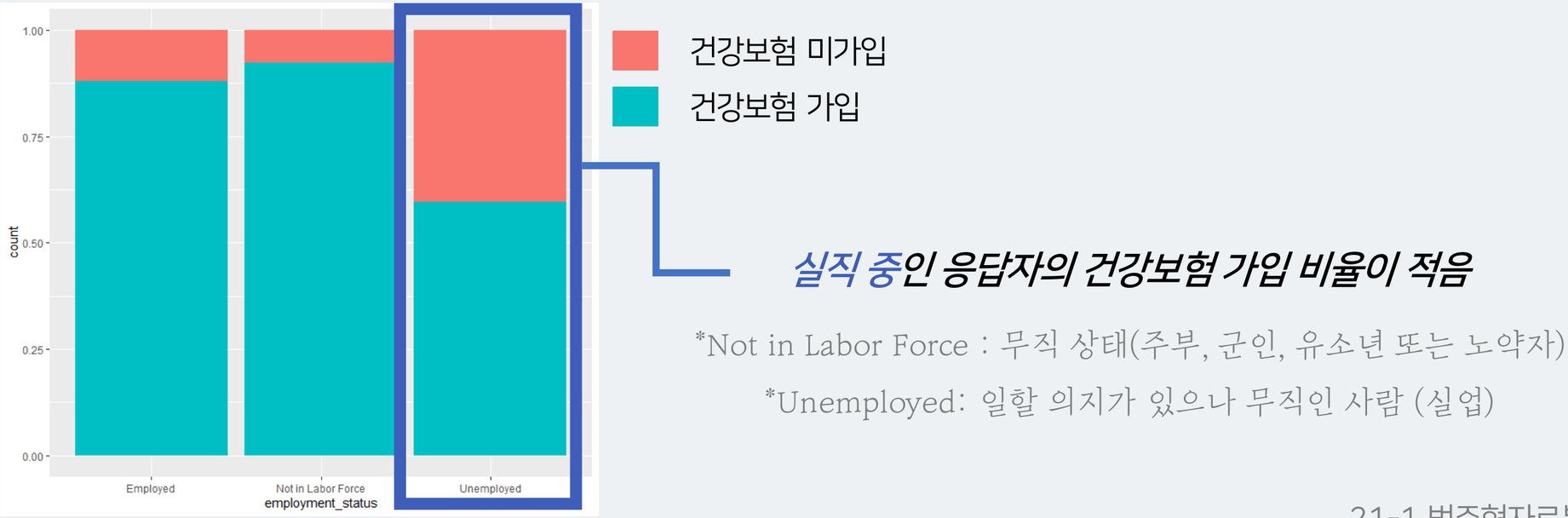
결측 비율이 가장 높았던 소득 변수와 건강보험유무 변수를 합리적 접근법을 활용하여 대체

IDEA

고용 상태 변수를 통해 **건강보험유무 변수**를 ,

자가 주택 소유 여부와 고용 상태 변수를 통해 **소득 변수**를 예측할 수 있지 않을까?

고용 상태별 건강보험 가입 여부

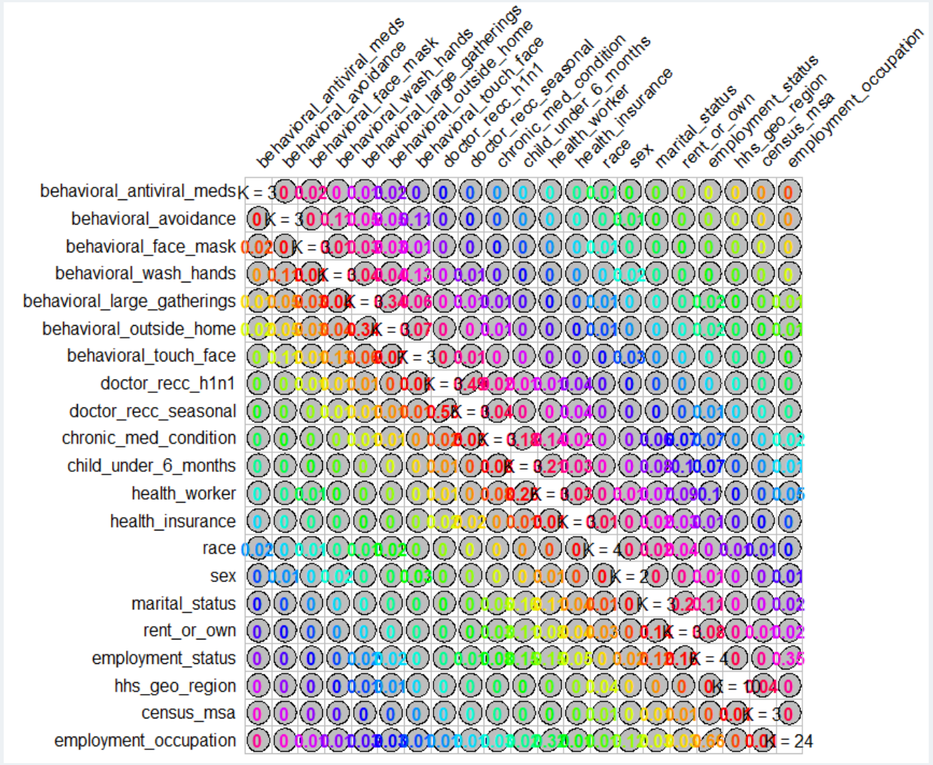


실직 중인 응답자의 건강보험 가입 비율이 적음

*Not in Labor Force : 무직 상태(주부, 군인, 유소년 또는 노약자)

*Unemployed: 일할 의지가 있으나 무직인 사람 (실업)

▪ 명목형 X변수 간의 연관성



doctor_recc_seasonal (계절독감 백신을 의사가 권장)

doctor_recc_h1n1(H1N1 백신을 의사가 권장)

: **Gktau 값 = 0.53**

behavioral_large_gatherings

(대규모 모임에서 시간 단축)

behavioral_outside_home

(외부인과 접촉을 줄임)

: **Gktau 값 = 0.34**

사회과학분야 연구니까 봐주떼염,,〉〈

~보노보노 피피티가
생각나는 그래프~
범주형 자료 ,, 쉽지 않다…

얘네 외에는 그닥,,,

▪ 순서형 X변수 간의 연관성

Spearman's Rank Correlation Coefficient

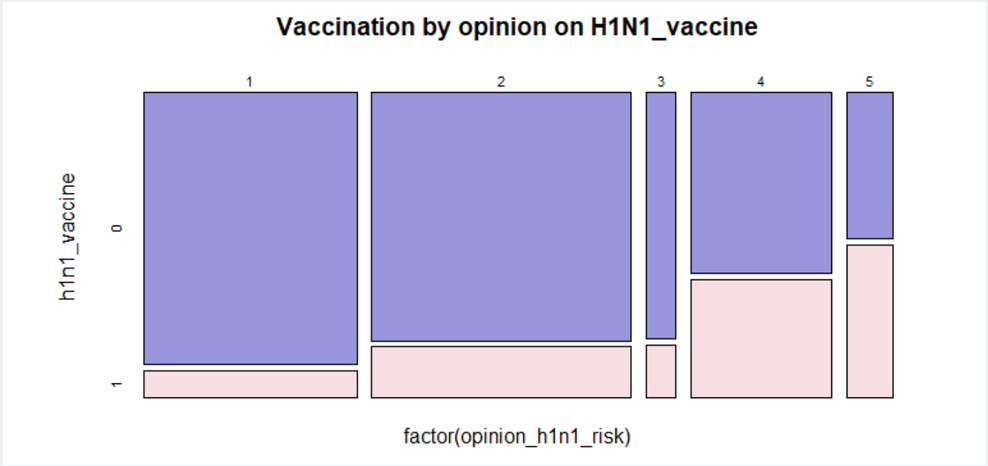
: Spearman 상관계수는 순서형 변수에 대한 연관성을 측정하는 측도

$$r_s = \frac{\sum (x' - \bar{x'})(y' - \bar{y'})}{\sqrt{\sum (x' - \bar{x'})^2 \sum (y' - \bar{y'})^2}}$$

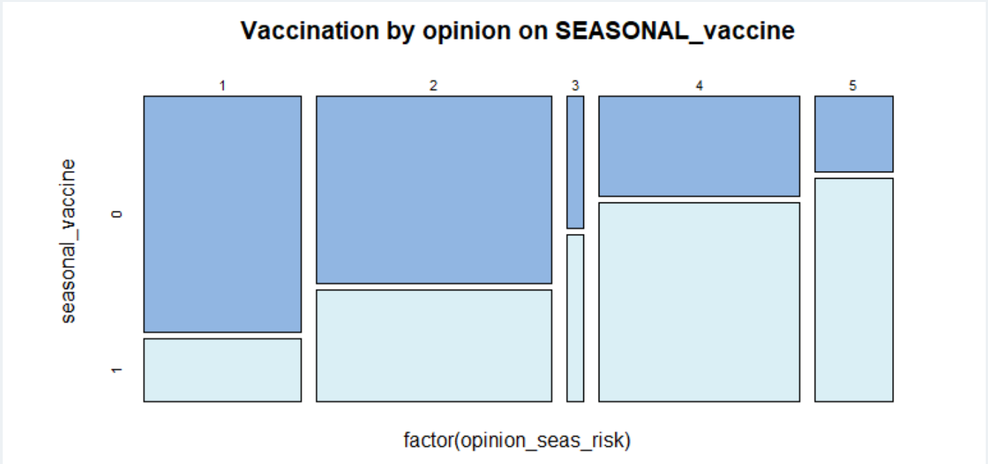
- x' 는 x의 순위, y' 는 y의 순위
- R에서 cor.test(method='spearman') 함수로 구할 수 있음

초딩 장히가 열심히 키우던 스피어맨,,

▪ 질병 위험에 대한 의견 (opinion)



1: 매우 낮음 5: 매우 높음



두 백신 접종 모두에서,

백신을 접종하지 않은 채 질병에

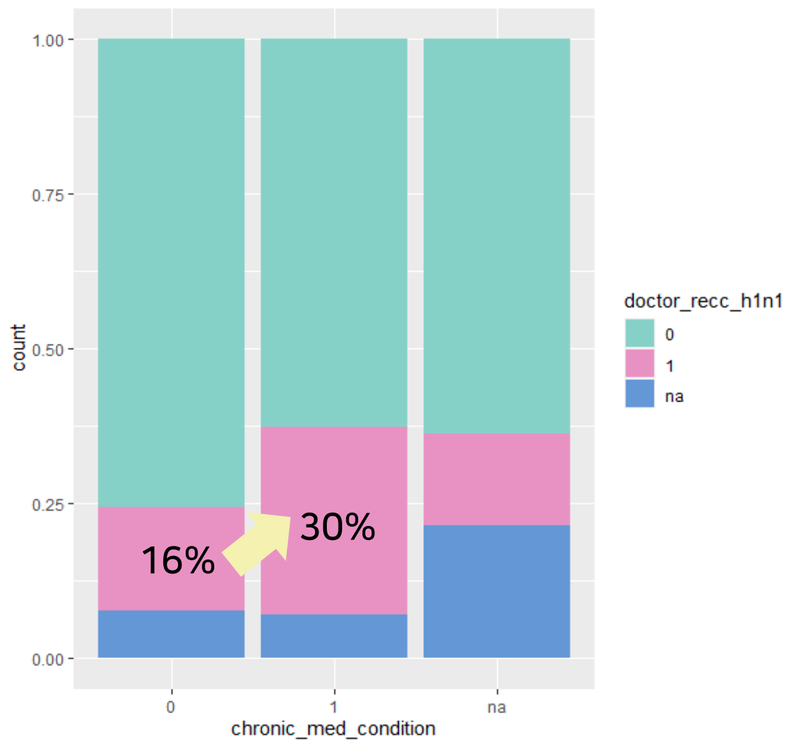
걸렸을 때 위험이 크다고 생각할수록

실제 백신 접종 비율이 높음

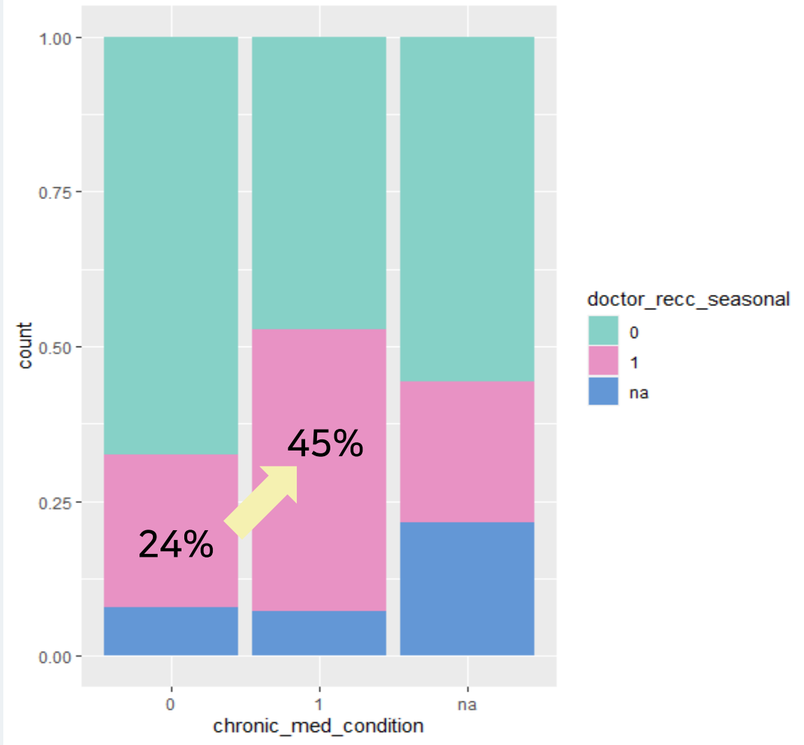
백신 없이 질병에 걸리는 것을
두려워할수록,
백신 접종할 확률이 높겠구나!

▪ 주변 의료적 환경

〈만성질환이 있는 환자에게
의사가 H1N1백신 권유했는지 여부 비율〉



〈만성질환이 있는 환자에게 의사가
계절독감 백신 권유했는지 여부 비율〉

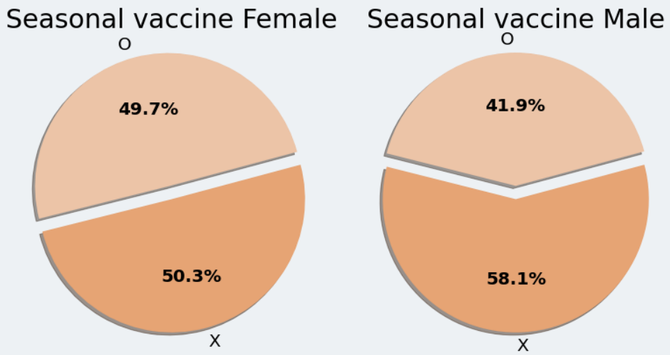


역시나 역할을 다하고 계시는 의사선생님 !

▪ 사회적, 경제적, 인구학적 배경

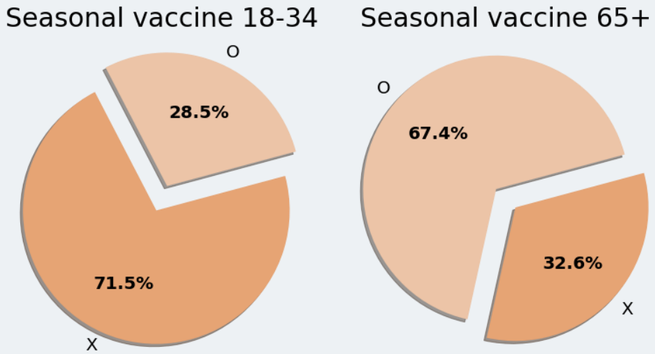
인적정보는 H1N1 백신여부에 별로 크게 영향을 끼치지 않았으므로
계절독감 백신여부 위주로 살펴보자 !

성별



여성이 남성보다 계절 독감 백신 접종률이 더 높음

나이



고연령층 사람들이 청년층보다
계절 독감 백신 접종률이 더 높음

▪ 건강신념모형(Helath Belief Model)이란?

건강 서비스 채택과 관련하여 건강 관련 행동을 설명하고 예측하기 위해 개발된 사회적, 심리적 모델

▪ 건강신념변인

지각된 심각성

질병 자체에 대해
심각하게 생각하는 정도

행동의 계기

질병 예방을 위해 필요하다고 여겨지는
적절한 행동을 촉발하는 요소

지각된 장애성

백신 접종에 대해
개인이 지각하고 있는 부정적인 요소

지각된 유익성

백신을 접종했을 때
얻을 수 있는 혜택에 대해 생각하는 정도

▪ 건강 신념 변인을 이용한 회귀식

$$(\text{h1n1\_vaccine}) = \beta_0 + \beta_1 \times (opinion\_\text{h1n1\_risk}) + \beta_2 \times (opinion\_h1n1\_vacc\_effective) + \beta_3 \times (opinion\_h1n1\_sick\_from\_vacc) + \beta_4 \times (doctor\_recc\_h1n1) + \beta_5 \times (h1n1\_knowledge) + \beta_6 \times (health\_worker)$$

$$(\text{seasonal\_vaccine}) = \beta_0 + \beta_1 \times (opinion\_\text{seas\_risk}) + \beta_2 \times (opinion\_seas\_vacc\_effective) + \beta_3 \times (opinion\_seas\_sick\_from\_vacc) + \beta_4 \times (doctor\_recc\_seasonal) + \beta_5 \times (health\_worker)$$

```
> anova(h1n1_glm, test = 'Chisq')
Analysis of Deviance Table

Model: binomial, link: logit

Response: h1n1_vaccine

Terms added sequentially (first to last)


                             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                         26706      27625
opinion_h1n1_risk             5  2692.26     26701      24933 < 2.2e-16 ***
opinion_h1n1_vacc_effective   5  1395.76     26696      23537 < 2.2e-16 ***
opinion_h1n1_sick_from_vacc   5   116.07     26691      23421 < 2.2e-16 ***
doctor_recc_h1n1              2  2261.08     26689      21160 < 2.2e-16 ***
h1n1_knowledge                3    87.26     26686      21073 < 2.2e-16 ***
health_worker                 2   253.08     26684      20820 < 2.2e-16 ***
---
```

```
> anova(seasonal_glm, test = 'Chisq')
Analysis of Deviance Table

Model: binomial, link: logit

Response: seasonal_vaccine

Terms added sequentially (first to last)


                             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                         26706      36897
opinion_seas_risk             5   4210.8     26701      32687 < 2.2e-16 ***
opinion_seas_vacc_effective   5   2438.2     26696      30248 < 2.2e-16 ***
opinion_seas_sick_from_vacc   5    552.8     26691      29696 < 2.2e-16 ***
doctor_recc_seasonal          2   2075.4     26689      27620 < 2.2e-16 ***
health_worker                 2    239.5     26687      27381 < 2.2e-16 ***
---
```

두 회귀식을 통해 건강신념모형의 유의성을 확인!

4주차 예고

~ 지연이는 2개의 종속변수를 예측하고 싶어~

Multi-Label Classification (다중 라벨 분류)

- *Algorithm adaptation methods*
  - 다변량 랜덤 포레스트
  - randomFerns
- *Problem transformation methods*
  - Binary Relevance
  - Chain Classifier 등등..

To be continued…

물어보지 마세요… 다음주의 범주팀에게…