ESC Final 22-Fall
Drug Consumption

이종현 | 김수민 김예찬 오재욱 윤이경 윤현석 천휴정 최가윤

목차

CHAPTER 01

from BIC to Bayes Factor

STEP 01

BIC

STEP 02

EDA

STEP 03

rstanarm

STEP 04

Result

1. Model Selection

1. AIC (Akaike Information Criterion)
2. BIC (Bayesian Information Criterion)

$$AIC_M = -2log\,L_M + 2p_M$$
$$BIC_M = -2log\,L_M + (log\,n) * p_M$$

- log L_M: model M의 log-likelihood
- p_M: Penalty Term. model M의 parameter 개수

$$AIC_M = -2log\,L_M + 2p_M$$
$$BIC_M = -2log\,L_M + (log\,n) * p_M$$

1. AIC 혹은 BIC의 값이 낮을수록 "상대적으로 좋은" 모델이다.

2. model M이 복잡해질수록(= parameter의 개수 증가) AIC 혹은 BIC의 값이 높아져 모델이 "좋지 않게" 된다. [Law of Parsimony]

3. log n ≥ 2 일 때, AIC에 비해 BIC의 penalty term 효과가 커진다.

2. Why Bayesian?

"n이 충분히 클 때" 두 모델 M1과 M2의 BIC를 비교하면,
두 모델 사이의 "Bayes Factor"인 bayesian model selection 기준을 approximate할 수 있다.

approximation 공식

$$BF_{12} \approx exp(\frac{BIC_1}{-2} - \frac{BIC_2}{-2})$$

3. Bayesian Factor

BF_{12} 공식

$$\frac{P(M_1|\mathbf{Y})}{P(M_2|\mathbf{Y})} = \frac{P(\mathbf{Y}|M_1)}{P(\mathbf{Y}|M_2)} * \frac{P(M_1)}{P(M_2)}$$

$P(M_1)$과 $P(M_2)$는 model 자체에 대한 prior로, parameter에 prior와 혼동해서는 안 된다. 예를 들어, 우리의 사전 지식에 따라 두 모델이 비슷하게 보인다면 P(M1) = P(M2) = 0.5과 같이 assign할 수 있다.

3. Bayesian Factor

$$\frac{P(M_1|\mathbf{Y})}{P(M_2|\mathbf{Y})} = \frac{P(\mathbf{Y}|M_1)}{P(\mathbf{Y}|M_2)} * \frac{P(M_1)}{P(M_2)}$$

1. Bayesian Factor: M_1과 M_2의 Posterior Model Probabillity인 $P(M_1|Y)$과 $P(M_2|Y)$ 사이의 비율. 이 비율 값이 클수록 M_1에게 유리하게 작용한다.

2. Rules of Thumb: 비율 값이 100이상이면 M_2 대신 M_1을 선택해야 한다는 강력한 증거.

3. Bayesian Factor

Typically, we will deal with parametric models M_k, which are described through model parameters, such as $\boldsymbol{\theta}_k$. So the marginal likelihoods $P(\boldsymbol{Y}|M_k)$ are evaluated using

$$P(\boldsymbol{Y}|M_k) = \int P(\boldsymbol{Y}|\boldsymbol{\theta}_k, M_k)P(\theta_k|\, M_k)d\theta_k, \tag{2}$$

where $P(\boldsymbol{Y}|\theta_k, M_k)$ is the likelihood under M_k, and $P(\theta_k|M_k)$ is the prior distribution of θ_k. A closed form analytical expression of the marginal likelihoods is difficult to obtain, even with completely specified priors (unless they are conjugate priors). Moreover, in most cases $\boldsymbol{\theta}_k$ will be high-dimensional and a direct numerical integration will be computationally intensive, if not impossible. For this reason we turn to a way to approximate this integral given in (2).

실제 상황에서 conjugacy 등이 성립하지 않는 이상 위 적분을 analytical하게 구할 수 있는 경우가 많지 않아서, Bayes Factor의 정확한 값을 구하기 어렵다.

4. Why Utilize BF?

Bayesian Model Averaging

여러 candidate model이 있을 때, 그 중 하나만을 선택하는 것이 아니라 각 모델에게 가중치를 부여해 모든 모델을 종합적으로 고려하는 "averaging". 일종의 Bayesian ensemble을 가능하게 한다

각 모델에게 부여되는 weight는 BF의 분자와 분모를 이뤘던 Posterior Model Probability (PMP) 이다.

K개의 후보 모델이 있을 때 k^{th} 모델의 PMP

$$p(M_k|\boldsymbol{Y}) = \frac{p(\boldsymbol{Y}|M_k)p(M_k)}{\sum_{k=1}^{K} p(\boldsymbol{Y}|M_k)p(M_k)}$$

$$p(\boldsymbol{Y}|M_k) = \int p(\boldsymbol{Y}|\theta_k, M_k)p(\theta_k|M_k)d\theta_k$$

이며, 이때 활용가능한 approximation이

$$p(\boldsymbol{Y}|M_k) = exp(\frac{BIC_k}{-2}) \quad (\text{assuming n is "large"})$$

4. Why Utilize BF?

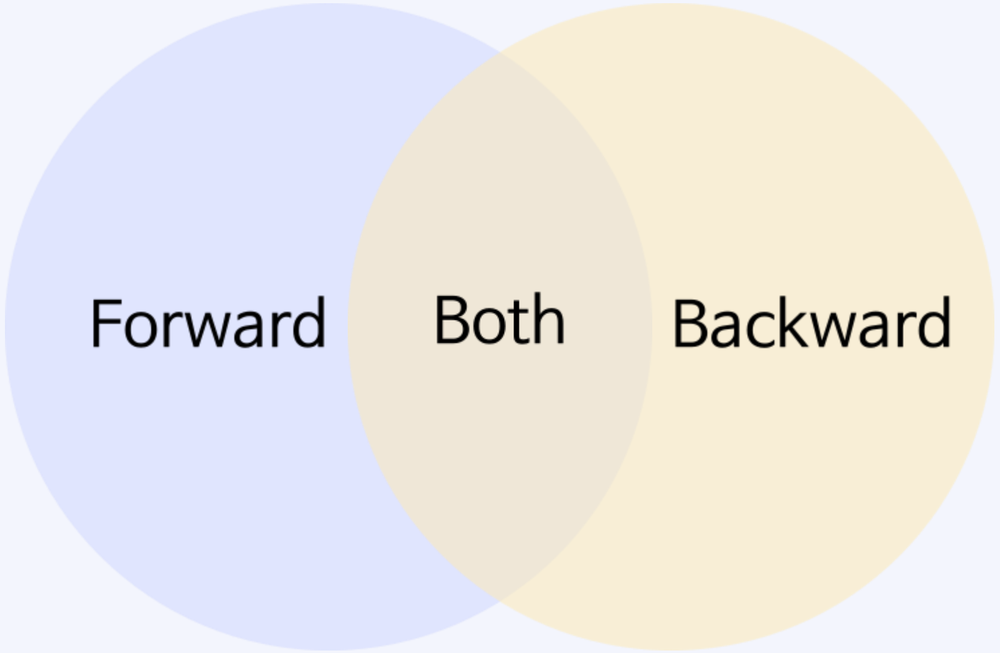
Then the BMA estimate of prediction at unobserved $\boldsymbol{Y}^*$ is

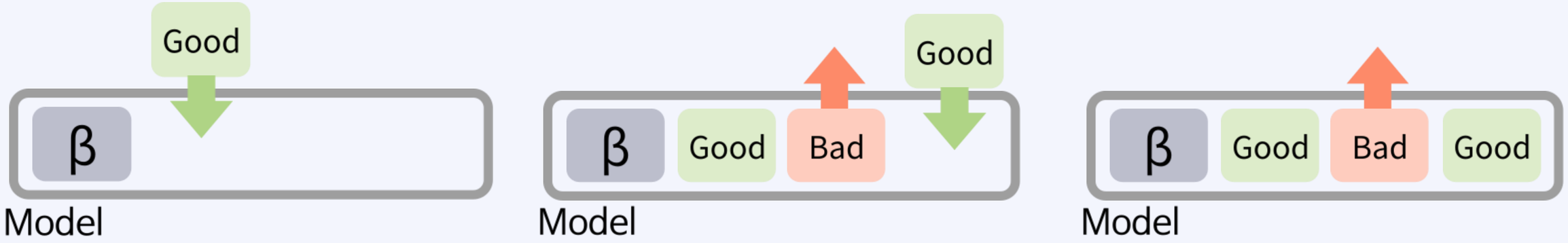
$$\boldsymbol{Y}^* = \sum_{k=1}^{K} \boldsymbol{Y}_k^* p(\mathcal{M}_k | \boldsymbol{Y})$$

where $\boldsymbol{Y}_k^*$ predicted response from $\mathcal{M}_k$

이처럼 BMA 앙상블 기법을 활용해 관측되지 않는 Y^* 값의 prediction이 가능하다

5. R 실습


Forward
Both
Backward


Good
β
Model
Good
β
Good
Bad
Model
Good
β
Good
Bad
Good
Model

step: Choose a model by AIC in a Stepwise Algorithm 함수

R Code

step(object, scope, scale = 0,

direction = c("both", "backward", "forward"),

trace = 1, keep = NULL, steps = 1000, k = 2, ...)

```
null=glm(Coke~1, data=overdose)
full=glm(Coke~., data=overdose)
step(null, scope=list(lower=null, upper=full), direction="forward", k=log(nrow(overdose)))
step(full, data=overdose, direction="backward", k=log(nrow(overdose)))
step(null, scope = list(upper=full), data=overdose, direction="both", k=log(nrow(overdose)))
```

https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/step

5. R 실습

Step1 Step2 Step3 Step4

```
Start:  AIC=2601.9
Coke ~ 1

             Df Deviance    AIC
+ Ectasy      1   261.48 1643.0
+ Amphet      1   289.82 1836.9
+ Mushrooms   1   335.98 2115.6
+ Cannabis    1   347.13 2177.1
+ Ketamine    1   348.54 2184.7
+ legath      1   350.01 2192.7
+ LSD         1   351.28 2199.5
+ Benzos      1   356.71 2228.4
+ Heroin      1   361.69 2254.6
+ Crack       1   368.17 2288.0
+ Amyl        1   373.36 2314.4
+ Nicotine    1   380.16 2348.4
+ Meth        1   382.95 2362.2
+ SS          1   388.67 2390.1
+ VSA         1   403.18 2459.2
+ Country     1   404.93 2467.4
+ Impulsive   1   408.96 2486.1
+ Age         1   413.08 2505.0
+ Oscore      1   419.80 2535.4
+ Cscore      1   419.94 2536.0
+ Ascore      1   423.64 2552.6
+ Gender      1   425.21 2559.5
+ Nscore      1   429.04 2576.4
+ Education   1   430.79 2584.1
+ Alchol      1   431.24 2586.1
+ Caff        1   432.32 2590.8
+ Ethnicity   1   433.70 2596.8
<none>           436.62 2601.9
+ Semer       1   435.30 2603.7
+ Escore      1   436.37 2608.3
+ Choc        1   436.60 2609.4
```

```
Step:  AIC=1643.01
Coke ~ Ectasy

             Df Deviance    AIC
+ Amphet      1   233.87 1440.2
+ Heroin      1   237.38 1468.3
+ Crack       1   238.81 1479.6
+ Benzos      1   243.91 1519.4
+ Meth        1   250.46 1569.4
+ Ketamine    1   252.22 1582.5
+ Amyl        1   252.42 1584.1
+ Cannabis    1   252.47 1584.5
+ Nicotine    1   253.53 1592.4
+ Mushrooms   1   254.56 1600.0
+ VSA         1   257.31 1620.2
+ Impulsive   1   257.50 1621.6
+ Ascore      1   257.50 1621.6
+ legath      1   257.82 1624.0
+ SS          1   258.28 1627.3
+ LSD         1   258.79 1631.0
+ Nscore      1   259.34 1635.0
+ Country     1   259.90 1639.1
+ Cscore      1   260.13 1640.8
<none>           261.48 1643.0
+ Alchol      1   260.61 1644.2
+ Ethnicity   1   260.78 1645.5
+ Semer       1   260.93 1646.6
+ Caff        1   261.05 1647.4
+ Choc        1   261.32 1649.4
+ Gender      1   261.39 1649.9
+ Age         1   261.43 1650.1
+ Oscore      1   261.43 1650.2
+ Escore      1   261.47 1650.5
+ Education   1   261.48 1650.5
```

5. R 실습

Step1 Step2 Step3 Step4

```
Step:  AIC=1440.17
Coke ~ Ectasy + Amphet

             Df Deviance    AIC
+ Heroin      1   219.57 1328.8
+ Crack       1   219.62 1329.2
+ Benzos      1   226.76 1389.6
+ Ketamine    1   228.76 1406.1
+ Amyl        1   229.62 1413.1
+ Nicotine    1   229.92 1415.6
+ Cannabis    1   230.48 1420.2
+ Meth        1   230.48 1420.2
+ Ascore      1   230.99 1424.3
+ Mushrooms   1   231.14 1425.6
+ Impulsive   1   232.35 1435.4
+ VSA         1   232.53 1436.9
+ Nscore      1   232.74 1438.6
+ SS          1   232.90 1439.9
<none>           233.87 1440.2
+ legath      1   233.11 1441.6
+ Alchol      1   233.13 1441.8
+ Age         1   233.19 1442.2
+ Semer       1   233.30 1443.1
+ Ethnicity   1   233.49 1444.7
+ LSD         1   233.54 1445.1
+ Cscore      1   233.55 1445.1
+ Education   1   233.66 1446.0
+ Caff        1   233.72 1446.5
+ Choc        1   233.76 1446.8
+ Gender      1   233.76 1446.8
+ Escore      1   233.82 1447.3
+ Oscore      1   233.85 1447.5
+ Country     1   233.85 1447.5
```

```
Step:  AIC=1255.54
Coke ~ Ectasy + Amphet + Heroin + Crack + Amyl

             Df Deviance    AIC
+ Nicotine    1   206.87 1239.1
+ Cannabis    1   207.18 1242.0
+ Benzos      1   207.46 1244.5
+ Ketamine    1   207.67 1246.4
+ Ascore      1   207.98 1249.2
+ Mushrooms   1   208.48 1253.8
<none>           209.52 1255.5
+ Alchol      1   208.87 1257.2
+ Impulsive   1   208.98 1258.2
+ legath      1   208.99 1258.3
+ Education   1   209.03 1258.7
+ Age         1   209.08 1259.2
+ Gender      1   209.10 1259.3
+ Nscore      1   209.12 1259.5
+ SS          1   209.12 1259.5
+ Semer       1   209.22 1260.4
+ Ethnicity   1   209.24 1260.5
+ Caff        1   209.34 1261.5
+ Escore      1   209.35 1261.5
+ Meth        1   209.40 1262.0
+ Cscore      1   209.43 1262.3
+ VSA         1   209.48 1262.8
+ Oscore      1   209.50 1262.9
+ Choc        1   209.50 1262.9
+ LSD         1   209.52 1263.1
+ Country     1   209.52 1263.1
```

5. R 실습


```
            Df Deviance    AIC
<none>           199.40 1215.0
+ Education  1   198.85 1217.3
+ Escore     1   198.86 1217.5
+ Mushrooms  1   198.96 1218.4
+ Alchol     1   199.06 1219.3
+ legath     1   199.12 1220.0
+ Semer      1   199.19 1220.6
+ Oscore     1   199.22 1220.8
+ Country    1   199.22 1220.8
+ SS         1   199.29 1221.5
+ Impulsive  1   199.30 1221.6
+ Ethnicity  1   199.31 1221.8
+ Caff       1   199.32 1221.8
+ Choc       1   199.35 1222.1
+ LSD        1   199.35 1222.1
+ Cscore     1   199.38 1222.3
+ VSA        1   199.38 1222.3
+ Nscore     1   199.40 1222.5
+ Meth       1   199.40 1222.5

Call:  glm(formula = Coke ~ Ectasy + Amphet + Heroin + Crack + Amyl +
    Nicotine + Benzos + Ketamine + Ascore + Gender + Age + Cannabis,
    data = overdose)

Coefficients:
(Intercept)       Ectasy       Amphet       Heroin        Crack         Amyl     Nicotine       Benzos     Ketamine       Ascore       Gender
   -0.02665      0.32438      0.18340      0.17659      0.18134      0.11249      0.06601      0.05523      0.09318     -0.02729      0.05605
        Age     Cannabis
    0.03529      0.07477

Degrees of Freedom: 1884 Total (i.e. Null);  1872 Residual
Null Deviance:      436.6
Residual Deviance: 199.4        AIC: 1143
```

CHAPTER 02

EDA와 데이터 전처리

STEP 01

BIC

STEP 02

EDA

STEP 03

rstanarm

STEP 04

Result

1. Semer

Step1 Step2 Step3 Step4

In [11]:
```
a=overdose[overdose['Semer']!='CL0'].index
overdose.drop(a)
```

Out[11]:

	ID	Age	Gender	Education	Country	Ethnicity	Nscore	Escore	Oscore	Ascore	...	Ectasy	Heroin	Ketamine	legath	LSD	Meth	Mushroo
0	1	0.49788	0	-0.05921	0.96082	0.12600	0.31287	-0.57545	-0.58331	-0.91699	...	CL0	CL0	CL0	CL0	CL0	CL0	
1	2	-0.07854	1	1.98437	0.96082	-0.31685	-0.67825	1.93886	1.43533	0.76096	...	CL4	CL0	CL2	CL0	CL2	CL3	
2	3	0.49788	1	-0.05921	0.96082	-0.31685	-0.46725	0.80523	-0.84732	-1.62090	...	CL0	CL0	CL0	CL0	CL0	CL0	
3	4	-0.95197	0	1.16365	0.96082	-0.31685	-0.14882	-0.80615	-0.01928	0.59042	...	CL0	CL0	CL2	CL0	CL0	CL0	
4	5	0.49788	0	1.98437	0.96082	-0.31685	0.73545	-1.63340	-0.45174	-0.30172	...	CL1	CL0	CL0	CL1	CL0	CL0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1880	1884	-0.95197	0	-0.61113	-0.57009	-0.31685	-1.19430	1.74091	1.88511	0.76096	...	CL0	CL0	CL0	CL3	CL3	CL0	
1881	1885	-0.95197	1	-0.61113	-0.57009	-0.31685	-0.24649	1.74091	0.58331	0.76096	...	CL2	CL0	CL0	CL3	CL5	CL4	
1882	1886	-0.07854	0	0.45468	-0.57009	-0.31685	1.13281	-1.37639	-1.27553	-1.77200	...	CL4	CL0	CL2	CL0	CL2	CL0	
1883	1887	-0.95197	0	-0.61113	-0.57009	-0.31685	0.91093	-1.92173	0.29338	-1.62090	...	CL3	CL0	CL0	CL3	CL3	CL0	
1884	1888	-0.95197	1	-0.61113	0.21128	-0.31685	-0.46725	2.12700	1.65653	1.11406	...	CL3	CL0	CL0	CL3	CL3	CL0	

2. 변수 변환

(1) 약물 레벨

In [12]:
```
addict_range={'CL0':0,'CL1':0,'CL2':1,'CL3':1,'CL4':1,'CL5':1,'CL6':1}
```

In [13]:
```
for i in range(13,32):
    overdose.iloc[:,i]=overdose.iloc[:,i].apply(lambda x : addict_range[x])
```

(2) 정규화된 변수 ⇒ nominal

In [16]:
```
education_range={'Education':{-2.43591:'Left school before 16yrs',
                              -1.73790:'Left school at 16yrs',
                              -1.43719:'Left school at 17yrs',
                              -1.22751:'Left school at 18yrs',
                              -0.61113:'Some college or university, no certificate/diploma',
                              -0.05921:'Professional certificate/diploma',
                               0.45468:'University degree',
                               1.16365:'Masters degree',
                               1.98437:'Doctorate degree'}}
```

3. Categorical Coding


```
overdose <- data
overdose$Country<-as.factor(overdose$Country)
overdose$Country
overdose$Country<-relevel(overdose$Country,ref='USA')
overdose$Ethnicity<-as.factor(overdose$Ethnicity)
overdose$Ethnicity<-relevel(overdose$Ethnicity,ref='Black')
overdose$Education<-as.factor(overdose$Education)
overdose$Eduation<-relevel(overdose$Education,ref='Doctorate degree')
overdose$Age<-as.factor(overdose$Age)
overdose$Age<-relevel(overdose$Age,ref='65+')
data <-overdose
data <- subset(data, select = -Semer)
```

3. Categorical Coding

Age, Gender, Education, Country, Ethinicity
: 숫자형으로 코딩된 각 변수의 value 값을 다시 매칭해서 범주형 변수로 변환

as.factor( )
: 특정 column을 vector를 기준으로 factorizing

Country
0.96082
0.96082
0.96082
0.96082
0.96082
0.24923
-0.57009
0.96082
0.24923
0.96082
0.96082
-0.28519
0.96082

Country
UK
UK
UK
UK
UK
Canada
USA
UK
Canada
UK
UK
Other
UK

4. Logistic Regression

로지스틱 회귀에서 범주형 변수의 의미

Example

(가정) country 변수에서 USA 를 기준으로 factorize

이후 특정 마약을 종속변수로 한 logistic regression 결과, Canada 변수의 기울기가 2라면
⇒ USA 에 비해서 Canada 국적을 가진 사람이 e^2배 더 해당 마약을 사용할 가능성이 있다
⇒ 해당 사건의 odd ratio가 e^2

Reference 설정 - relevel( )

Country	USA
Ethnicity	Black
Education	Doctorate degree
Age	65+

Ordinal

CHAPTER 03

rstanarm

STEP 01

BIC

STEP 02

EDA

STEP 03

rstanarm

STEP 04

Result

1. Why rstanarm?

glm function

: 일반화된 선형 모델의 최대 우도 추정 수행

rstanarm package

: stan_glm 함수를 사용해 binomial GLM
MCMC를 통한 완전한 bayesian 추정 수행

2. Distribution

prior 설정하기

R Code

```
t_prior <- student_t(df = 7, location = 0, scale = 2.5)
Caff_post <- stan_glm(formula, data = overdose,
                      family = binomial(link = "logit"),
                      prior = t_prior, prior_intercept = t_prior,
                      QR=TRUE, seed = 2001, refresh=0)
```

MCMC를 이용한 posterior distribution 도출

: posterior median estimates와 90% intervals을 구할 수 있다

CHAPTER 04

Result

STEP 01

BIC

STEP 02

EDA

STEP 03

rstanarm

STEP 04

Result

1. 경향성

Age	환각버섯, 니코틴, vsa, 케타민, 코카인, 리갈하이, lsd, 엑스터시 , 암페타민, 아밀, 대마 회귀계수의 신뢰구간이 0을 포함하거나 신뢰구간의 크기가 너무 커 결론을 내리기 어려움
Gender	리갈하이, lsd, meth, 엑스터시, 암페타민, 아밀, 크랙, 대마, 환각버섯, 니코틴, 케타민, 히로인 유의미한 확률로 남자일 때 해당 마약을 할 가능성 높음
Country	뉴질랜드: 리갈하이, 대마 | 오스트레일리아: 암페타민 | 영국: 아밀 회귀계수의 신뢰구간이 0을 포함하는 경우가 다수 있어 신뢰하기 어려움
Ethnicity	환각버섯 흑인보다 백인, 백인-아시아인 혼혈이 해당 마약을 할 가능성 높음

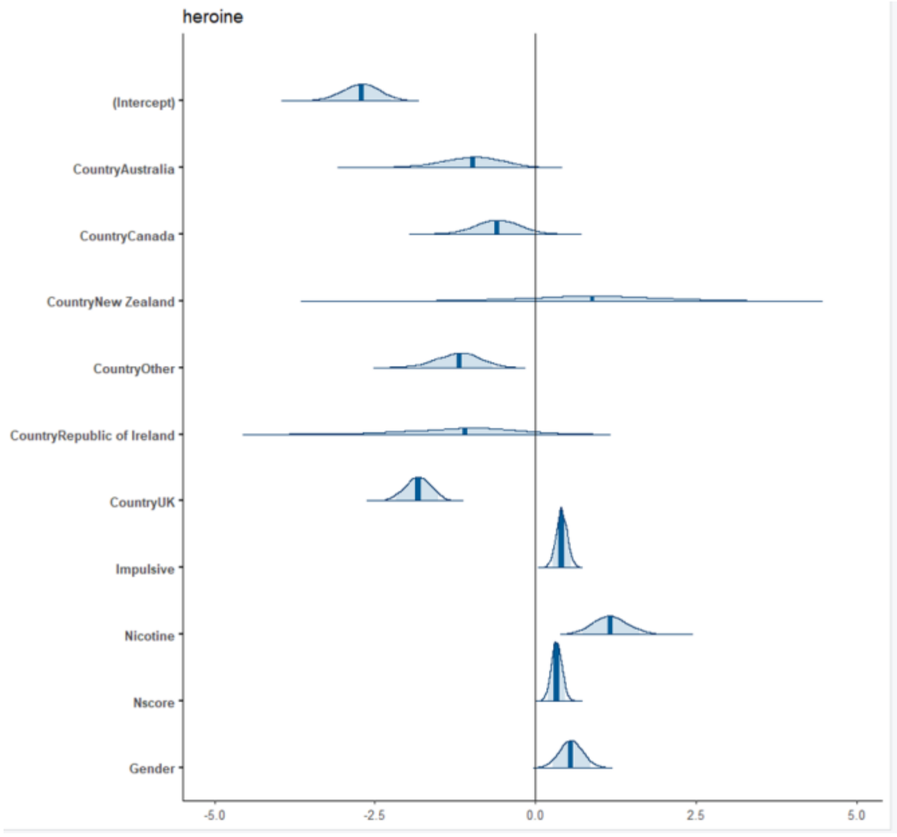
1. 경향성

Step1 Step2 Step3 Step4

Nscore	meth, 헤로인 신경증 점수가 높을수록 해당 마약을 할 가능성 높음
Oscore	리갈하이, LSD, 엑스터시, 벤조디아제핀, 대마, 환각버섯, 니코틴, 케타민 개방성 점수가 높을수록 해당 마약을 할 가능성 높음
Cscore	리갈하이, 엑스터시, 암페타민, 카나비스, 니코틴, VSA, 케타민, 코카인 모든 약물에 대해 성실성 점수가 낮을수록 해당 약물을 할 가능성 높음
Impulsize	크랙, 히로인 충동성 점수가 높을수록 해당 약물을 할 가능성 높음
SS	리갈하이, LSD, 액스터시 등 모든 약물에 대해서 감각추구 점수가 높을수록 마약을 할 가능성 높음

2. Bayesian Analysis

(1) 헤로인



	5%	95%
(Intercept)	-3.21	-2.25
CountryAustralia	-1.85	-0.28
CountryCanada	-1.18	-0.02
CountryNew Zealand	-0.85	2.63
CountryOther	-1.79	-0.67
CountryRepublic of Ireland	-2.66	0.10
CountryUK	-2.17	-1.50
Impulsive	0.26	0.55
Nicotine	0.74	1.63
Nscore	0.20	0.46
Gender	0.27	0.86

β 의 의미

- Logistic regression: $log(\frac{p(x)}{1-p(x)}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$
- Odds: $odds(x) = \frac{p(x)}{1-p(x)} = e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}$
- Odds ratio between x_k=1 and x_k=0 (fix others): $e^{p_k} = \frac{p(Y=1|x_k)}{p(Y=1|x_0)}$

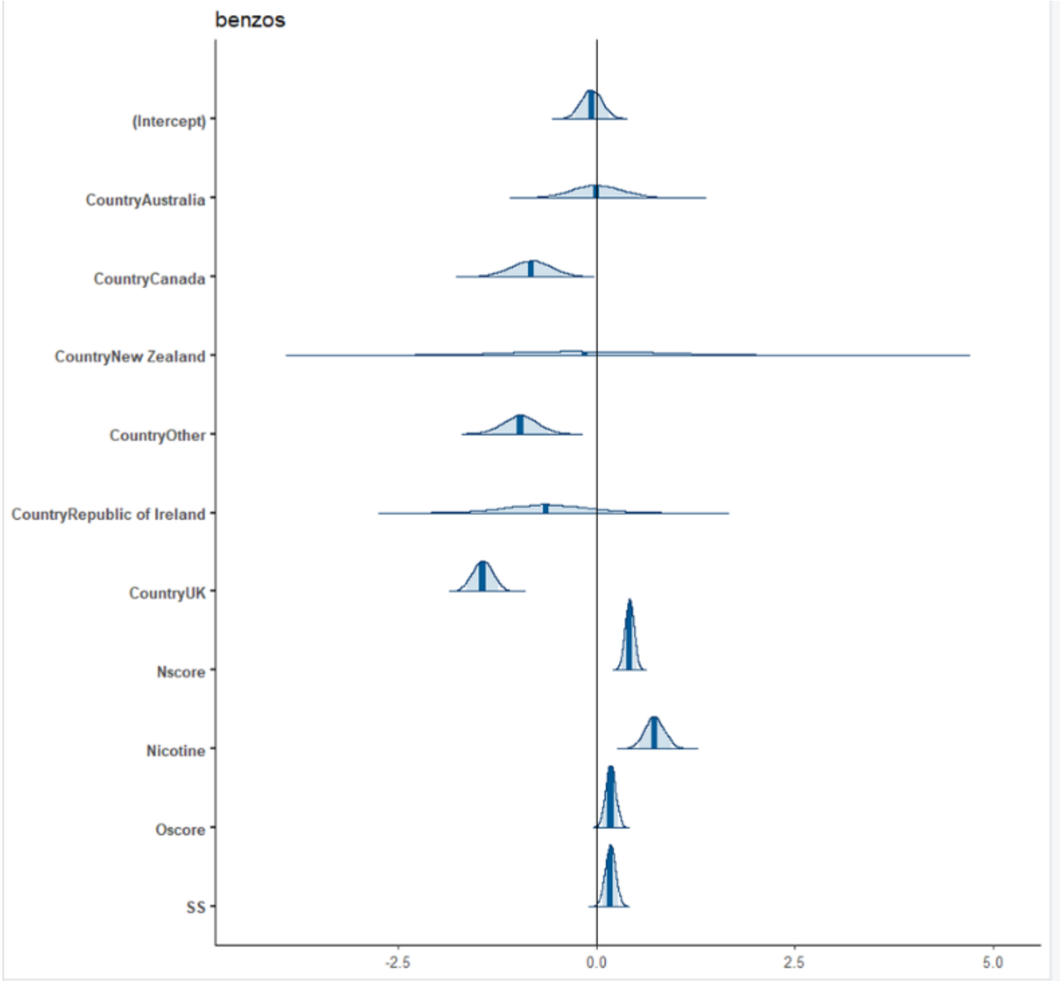
$$\frac{odds(x_{k=1})}{odds(x_{k=0})} = \frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \ldots + \beta_p x_p}}{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_k 0 + \ldots + \beta_p x_p}} = e^{\beta_k x_k} = e^{\beta_k}$$

- Interpretation of β_k:

  If we increase x_k 1, odds is increased by $e^{\beta k}$

  - β_k>0: p(Y=1 | xk) becomes increased as x_k large
  - β_k<0: p(Y=1 | xk) becomes decreased as x_k large

(2) 벤조디아제핀




```
> round(posterior_interval(benzos, prob = 0.9), 2)
                                5%   95%
(Intercept)                  -0.27  0.17
CountryAustralia             -0.50  0.52
CountryCanada                -1.24 -0.42
CountryNew Zealand           -1.75  1.63
CountryOther                 -1.32 -0.60
CountryRepublic of Ireland   -1.44  0.21
CountryUK                    -1.64 -1.23
Nscore                        0.33  0.51
Nicotine                      0.53  0.94
Oscore                        0.08  0.28
SS                            0.07  0.28
>
```


THANK
YOU