19-2 YONSEI ESC FINAL PROJECT

BAYESIAN LINEAR REGRESSION WITH CRIMEDATA.csv

26 NOV 2019

3조

조장 : 홍익선
조원 : 강경훈 남승지 박건우 이가은 정규형

Table of Contents

Final Project 2nd week

I. Intro

Prior to the beginning

역할분담

ESC

- 총 **네 번**의 만남, **전원출석**, 24시간의 running time, 환상의 팀워크

1. 홍익선 (조장) : Re-feature Engineering & 예약의 달인
2. 남승지 : Re-feature Engineering & 슈퍼컴퓨터
3. 강경훈 : Coding machine & 발표
4. 박건우 : Storyteller & 지역론자
5. 정규형 : Mapping & Gaussian
6. 이가은 : Mapping & HTML 장인

II. Data and Our Approach

What is crimedata.csv?

무엇을 담고 있는 자료인가?

- **Crimedata.csv는 무엇인가?**
  - 미국 각 주의 community별 범죄율과 관련 사회경제 지표를 종합한 다변량 자료
  - UCI Machine Learning Repository에서 머신 러닝 코드 연습용으로 제작
  - US Census(1990), US FBI Report(1995), US Law Enforcement Survey(1990) 등에서 취합했으며, 서로 다른 데이터 셋을 취합하는 과정에서 NA 결측치가 발생
  - **각 Predictor에 가중치를 부여하여 종속변수를 설명하는 회귀분석 모델을 시험할 수 있음**

[Fig.] 데이터 개요

Data Set	**Communities and Crime Unnormalized Data Set**		
N of Attributes	147	**N of Instances**	2215
Potential goal	8 (강력범죄 (살인, 강간, 강도, 폭행) 및 비강력범죄 (절도, 상해, 방화, 차량 탈취))		
Predictive attributes	125 (인구, 연령, 학력, 인종 별 비율, 중위 소득, 가구 수 등)		
Non-predictive	4 (Community 및 주 이름, 코드 등)		

What is crimedata.csv?

무엇을 담고 있는 자료인가?

• Model Assumptions

1. Assumptions for Bayesian Inference :

① **Normal Sampling Density**
범죄 건수 Y(arsons, rapes, …)는 설명변수 X가 주어졌을 때 다변수 정규분포를 따른다.

② **“g-prior” (ch. 09)**
데이터 (Y, X)가 주어졌을 때 σ^2는 인버스 감마분포를 따르며,
데이터와 σ^2이 주어졌을 때 $\boldsymbol{\beta}$는 g-prior 형태의 다변수 정규분포를 따른다.

2. Assumptions on predictor variables:

① **Homoscedasticity**
모든 지역에 걸쳐 오차항의 분산은 동일하다.

② **Fixed Slopes**
모든 지역에 걸쳐 설명변수 (예컨대 인구 밀도)가 대상 범죄율에 미치는 영향은 같다.
즉 베타 값이 지역에 따라 달라지지 않는다!

③ **Varying Intercepts**
데이터에 포함된 모든 설명변수의 값이 동등할 때 지역별 평균 범죄율이 다를 수 있다.
즉 ‘텍사스이기 때문에’ (데이터에 없는 모종의 이유로) 대상 범죄율이 높을 수 있다!

What is crimedata.csv?

무엇을 담고 있는 자료인가?

- **Overall Workflow**

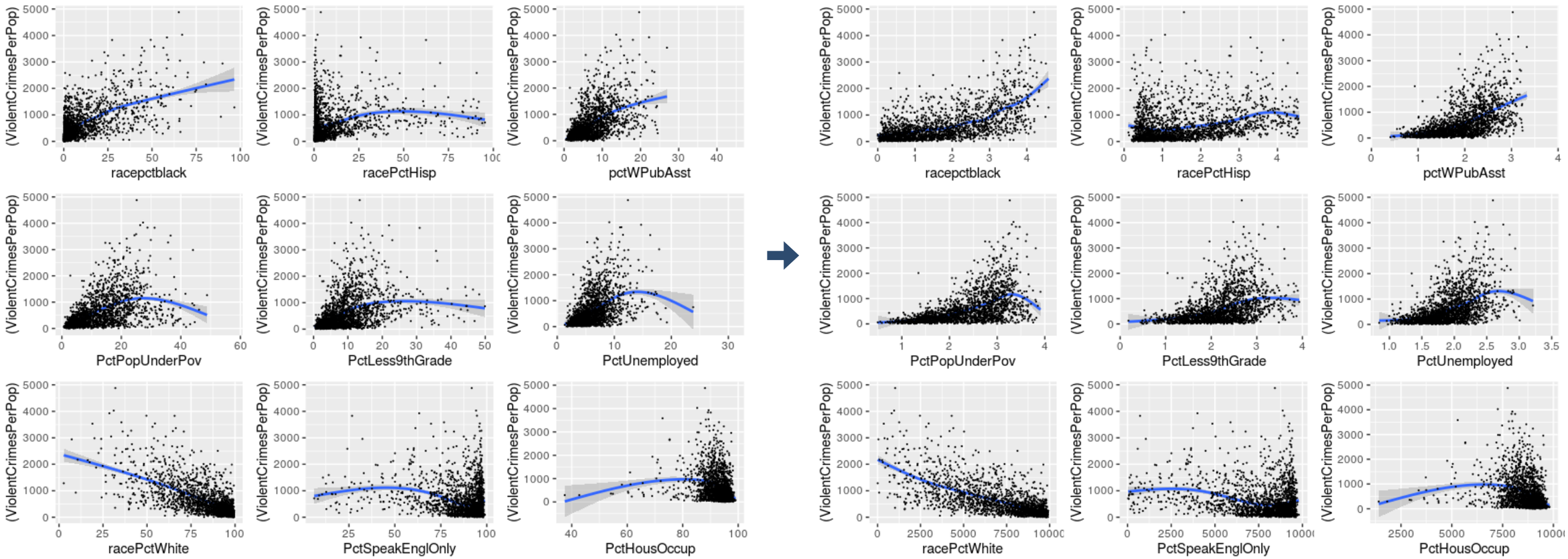
1. **Data Manipulation:**
   - ① NA imputation: Gibbs Sampler를 이용하여 설명변수 NA Imputation(ch. 07)
   - ② Log transformation: 종속변수는 Normal Sampling 가정을 고려하여 로그 변환, 설명변수는 종속변수와의 선형 관계를 위해 로그 변환
   - ③ Correlation control: 설명변수 간 상관계수가 0.9 이상인 변수 제거
   - ④ Feature scaling: 베타 계수 간 비교를 위해 모든 설명변수에 대해 Z-score scaling
   - ⑤ Stratified Sampling: 모델의 성능 비교를 위해서 지역별로 train-test 층화 추출
2. **Model Selection:**
   - ① 모델 선택에 대한 빈도론적 방법과 베이지안 방법의 비교
   - ② Z mean, Pred MSE, BIC 등을 고려하여 모델 (설명변수의 개수) 결정
   - ③ BMA와 OLS estimator의 성능 비교
3. **Model Description:**
   - ① 각 대상 범죄별로 총 8개의 Model
   - ② 어떤 설명변수가 어느 범죄에 기여하였나

III. Data Manipulation

Regression Assumption

가정에 충실하자!

- X's transformation : 총 23개의 X 변수를 Y(ViolentCrimesPerPop)과의 linearity를 위한 transformation. (logx,x^2변환)



Data Manipulation

Correlation control

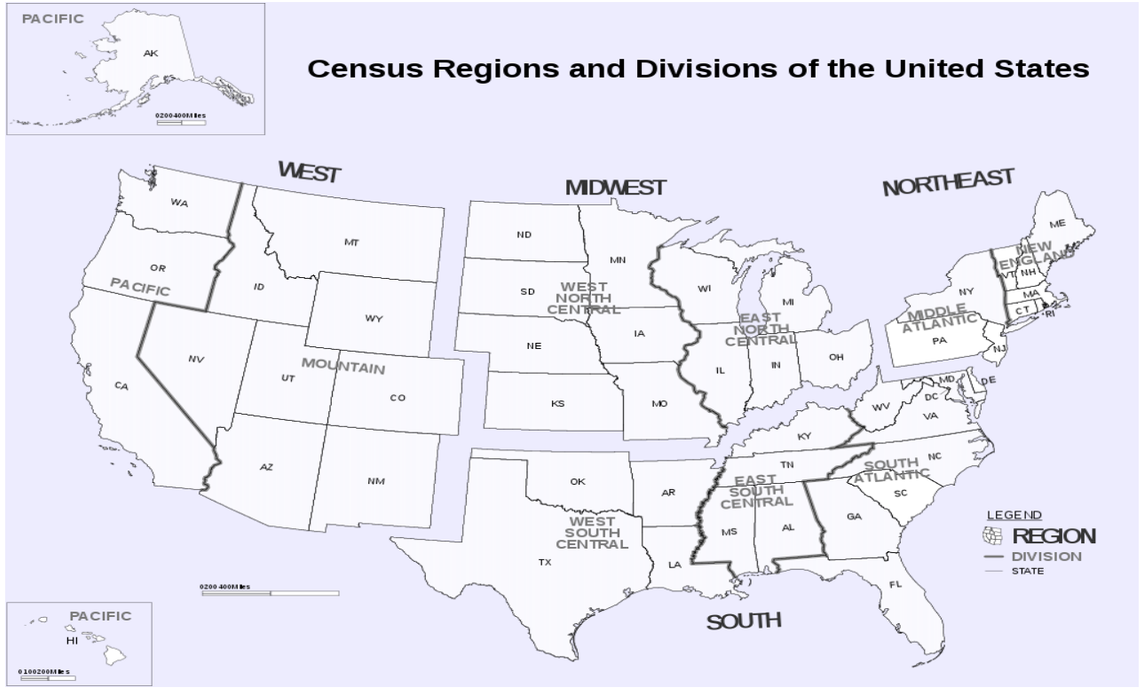
제거한 변수	그와 Corr이 높은 변수
NumInShelters	population
pctWSocSec	pctWWage, agePct65up
perCapInc	whitePerCap
PctNotHSGrad	PctLess9thGrade
PctOccupMgmtProf	PctBSorMore
PctLargHouseOccup	PersPerFam
PctRecImmig10	PctForeignBorn
PctNotSpeakEnglWell	PctSpeakEnglOnly
OwnOccLowQuart	OwnOccMedVal
OwnOccHiQuart	OwnOccMedVal
RentLowQ	MedRent
RentHighQ	MedRent
RentMedian	MedRent

IV. Model Selection : Frequentist vs Bayesian

Stratified Validation Set

The more information, the smaller variance

- Prediction set (with NA) / train set (70% of each division) / test set (30% of each division)
- Why stratification?
- For each state?



Variable for division

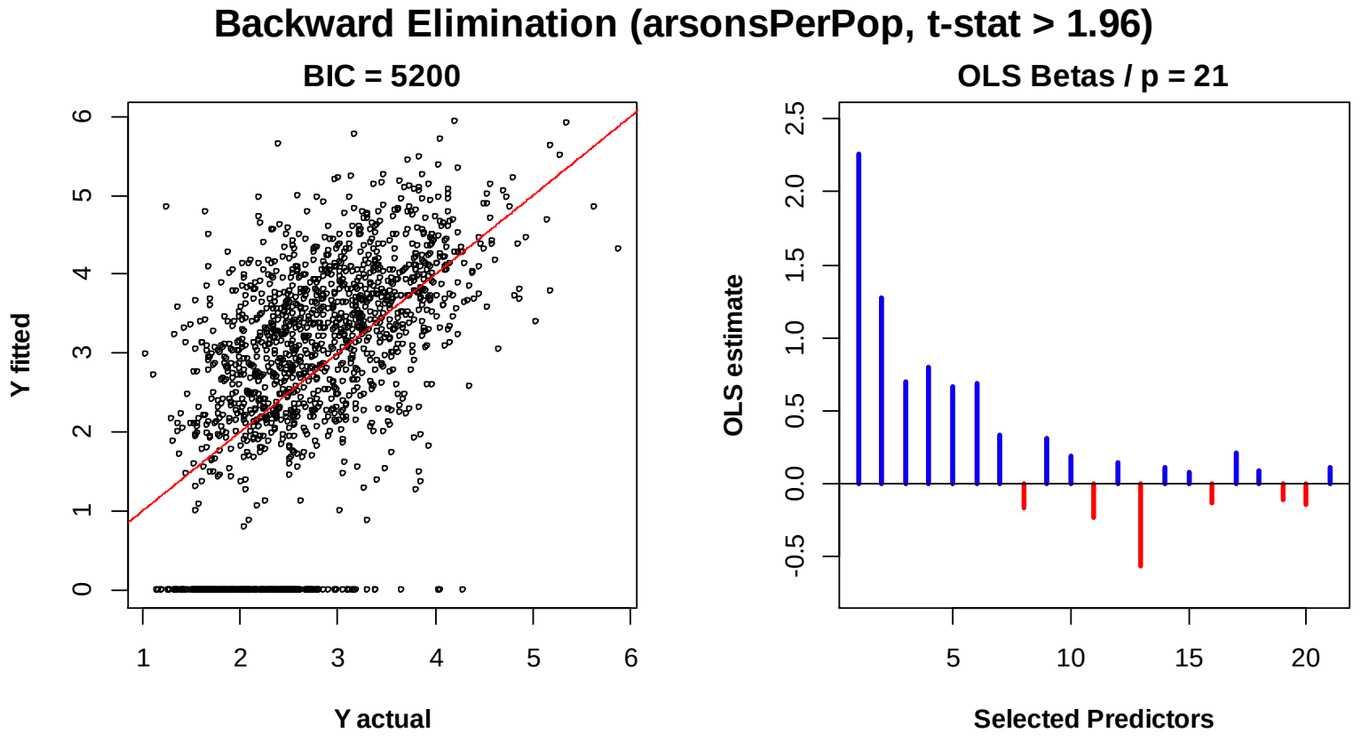
8 dummy variables.

- **Is it reasonable to include division variable in our model?**
  - For Omitted Variable
  - For Geographical Factor
  - For Gotham….

Model Selection

빈도론적 방법론과 한계

• Frequentist Approach for Model Selection: Stepwise Regression

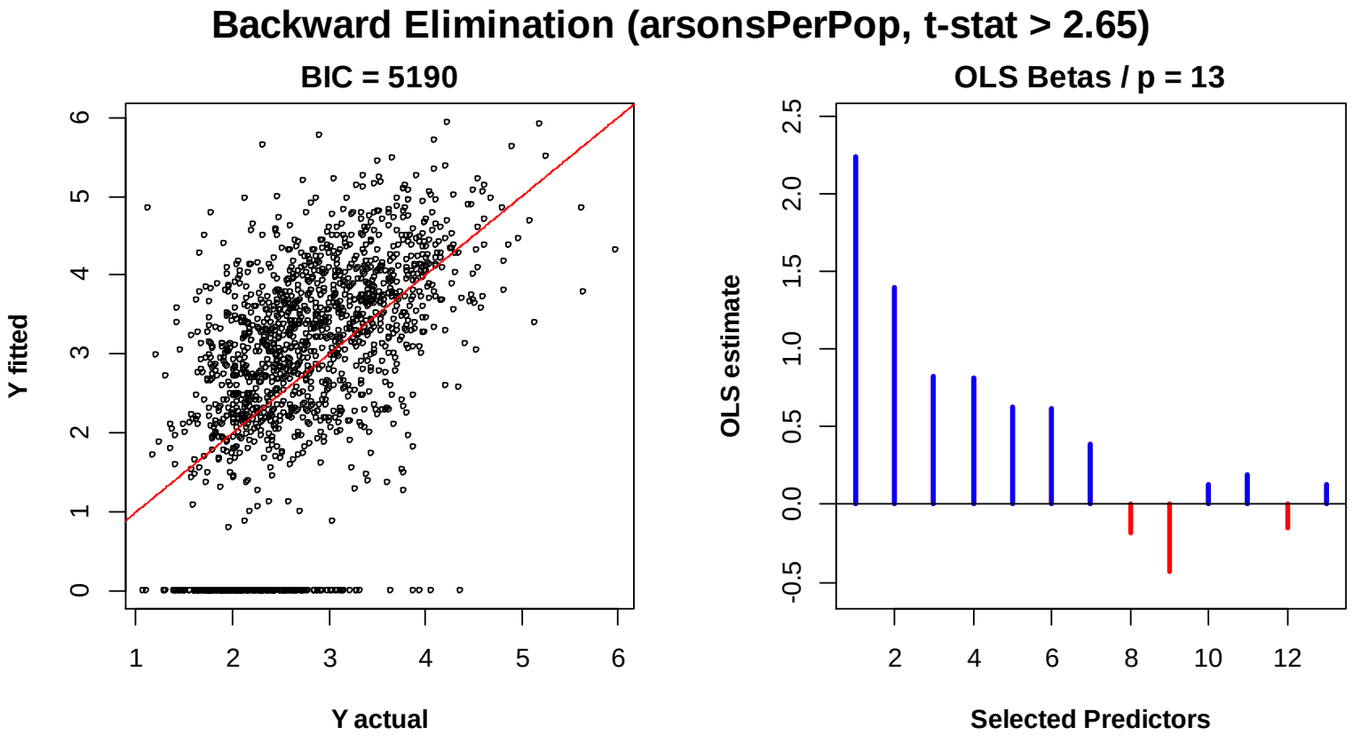


- 선형 회귀모형에서 모델의 선택은 곧 Predictor 개수를 결정하는 것
- 오차항의 정규분포를 가정하면 개별 베타의 검정에는 t-test, 여러 베타의 검정에는 F-test를 사용
- 개별 변수를 모델에 포함하고 제외하면서 기준 통계량 (예컨대 t-stat, BIC 등)의 변화를 보고 적정 변수를 선택하는 것이 Stepwise Regression
- 방향은 Backward, Forward, Bi-direction이 있으나 결과에 큰 차이는 없음.
- 여기에서는 Plot으로 나타내기 위해 Backward 방법을 사용하여 비교

Model Selection

빈도론적 방법론과 한계

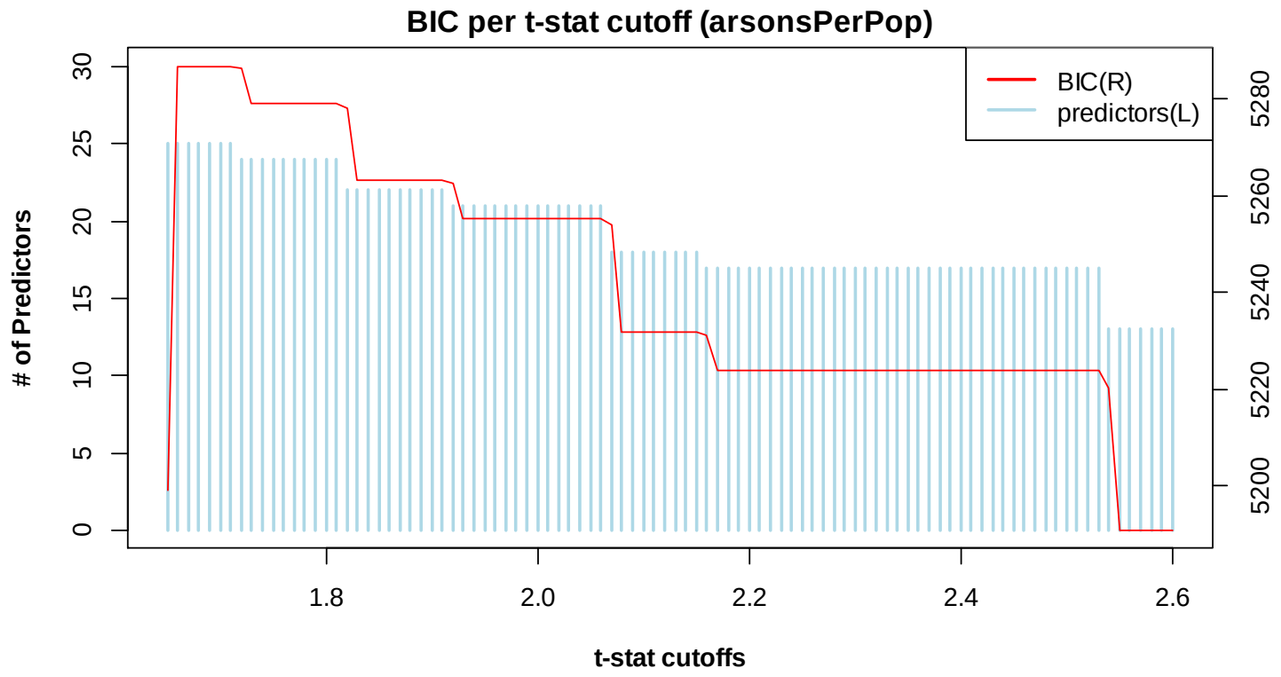
• Frequentist Approach for Model Selection: Stepwise Regression



- 선형 회귀모형에서 모델의 선택은 곧 Predictor 개수를 결정하는 것
- 오차항의 정규분포를 가정하면 개별 베타의 검정에는 t-test, 여러 베타의 검정에는 F-test를 사용
- 개별 변수를 모델에 포함하고 제외하면서 기준 통계량 (예컨대 t-stat, BIC 등)의 변화를 보고 적정 변수를 선택하는 것이 Stepwise Regression
- 방향은 Backward, Forward, Bidirection이 있으나 결과에 큰 차이는 없음.
- 여기에서는 Plot으로 나타내기 위해 Backward 방법을 사용하여 비교

Model Selection
빈도론적 방법론과 한계

• Frequentist Approach for Model Selection: Stepwise Regression



- 가장 최적의 모델은 가능한 한 적은 변수(parsimonious)로 가능한 한 많은 것을 설명할 수 있는(explainable) 모델
- 이 두 가지 조건을 모두 포함하여 <u>모델의 비교를 위해</u> 만든 통계량이 AIC, BIC

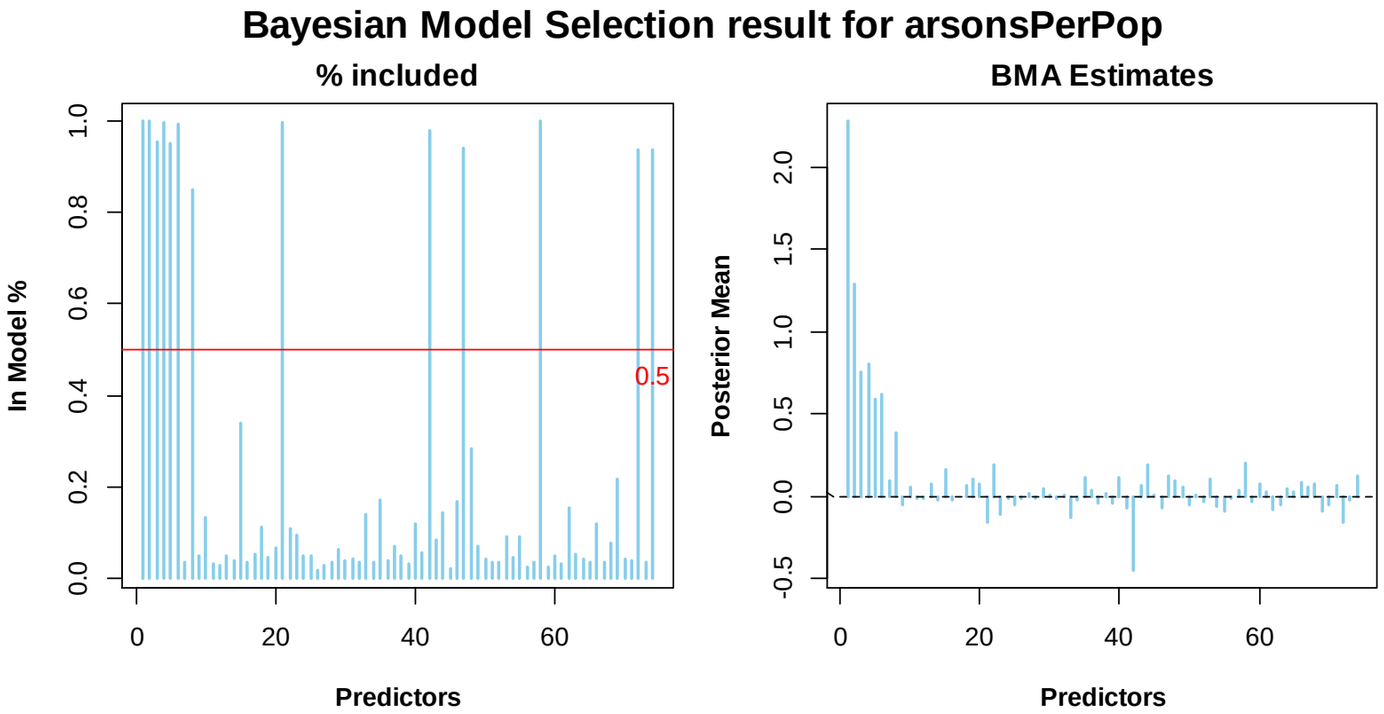
  AIC = −2loglikelihood + 2p
  BIC = −2loglikelihood + log(n)p

  Note: Bias를 변수를 제거하면서 생기는 오류, Var를 변수를 추가함으로서 생기는 오류로 본다면, AIC는 Bias와 Var의 합으로 볼 수 있다.

- AIC를 최소화하는 모델은 우도를 가장 크게 하면서 동시에 변수 개수가 가장 적은 모델

Model Selection
빈도론적 방법론과 한계

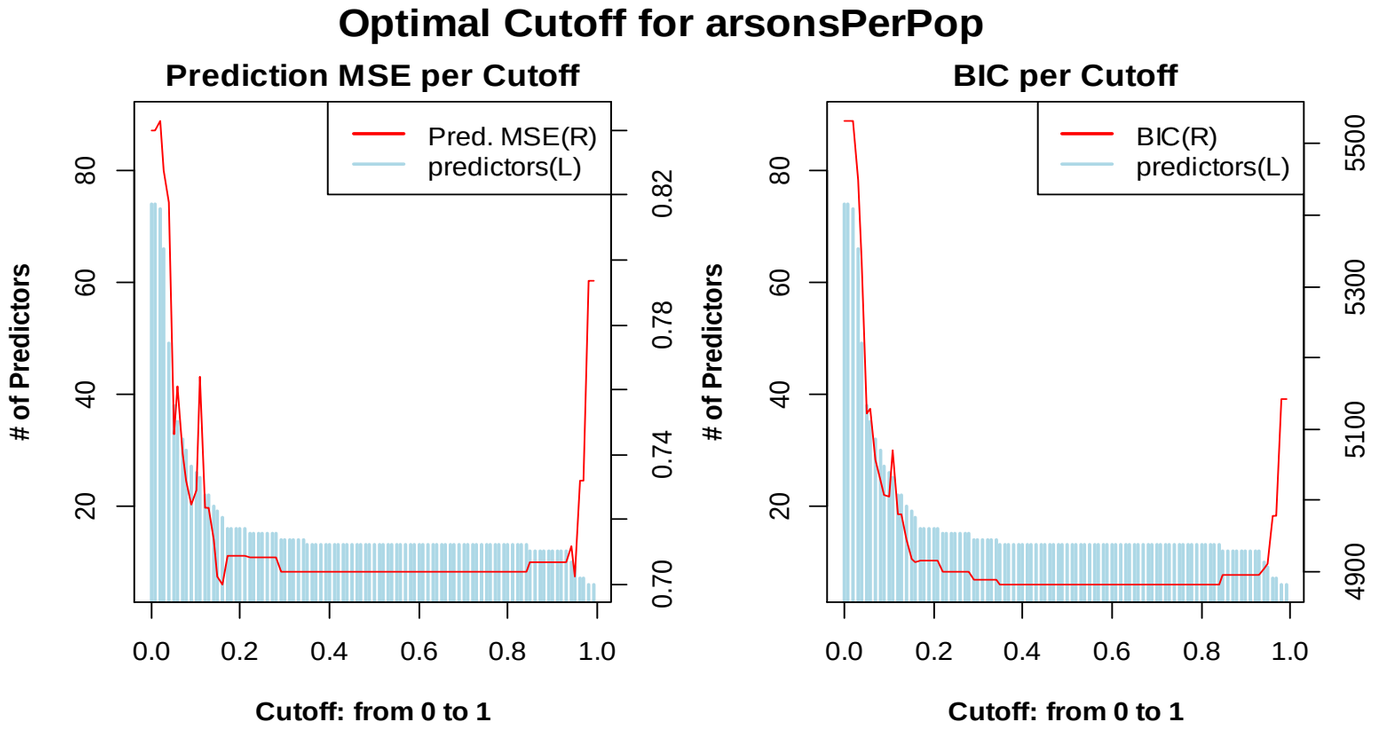
• Bayesian Approach for Model Selection: Model Likelihood



- 방법론 측면에서 베이지안 접근의 가장 큰 특징은 P[Z | X,Y], 즉 데이터가 주어졌을 때 모델의 posterior probability를 근사할 수 있다는 것.
- (Model Space에서의 균일 분포를 가정)
  이는 즉 P[Y |Xz(reduced dataset)], 모델의 Likelihood을 계산하는 것과 동일.
  G-prior를 사용하면 계산 가능!
- Z_mean:
  1000번의 깁스 샘플러 시행 중 각 predictor가 모델에 포함된 비율
  Zi의 marginal post. probability 평균의 추정치로 간주할 수 있다!

Model Selection
빈도론적 방법론과 한계

- Bayesian Approach for Model Selection: Model Likelihood

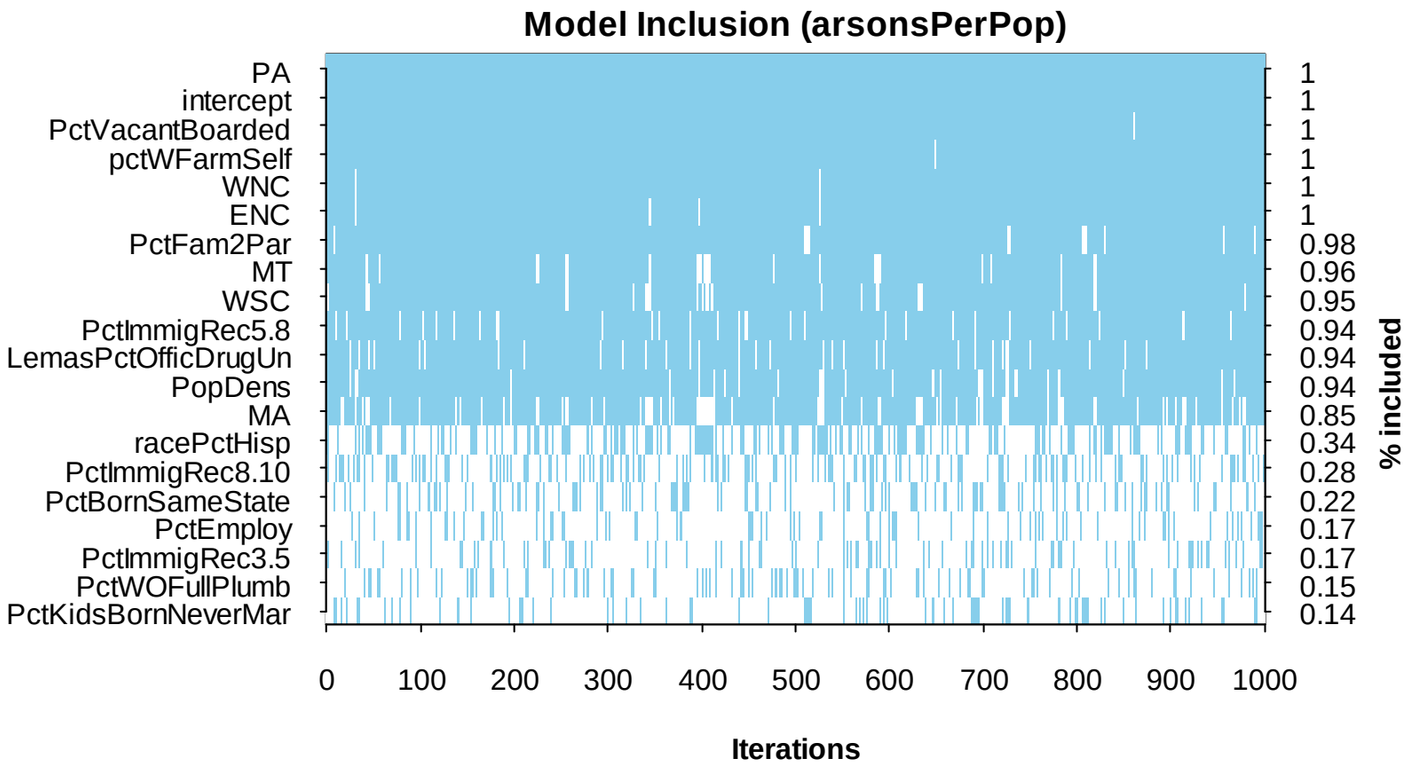


- How to Interpret Gibbs Sampler Result:
  - 깁스 샘플러로 생성한 Z_mean을 보고 predictor를 선정하는 것은 다소 작위적. 우리는 다음과 같은 판단 기준을 세움
  - 가능한 가장 적은 predictor 개수로
  ① Prediction MSE
  ② BIC
  를 최소화하는 Z-cutoff 선택
  - Model Iteration Plot를 보고 Z-cutoff 값과 각 변수의 Z_mean 비교

Model Selection

빈도론적 방법론과 한계

• Bayesian Approach for Model Selection: Model Likelihood

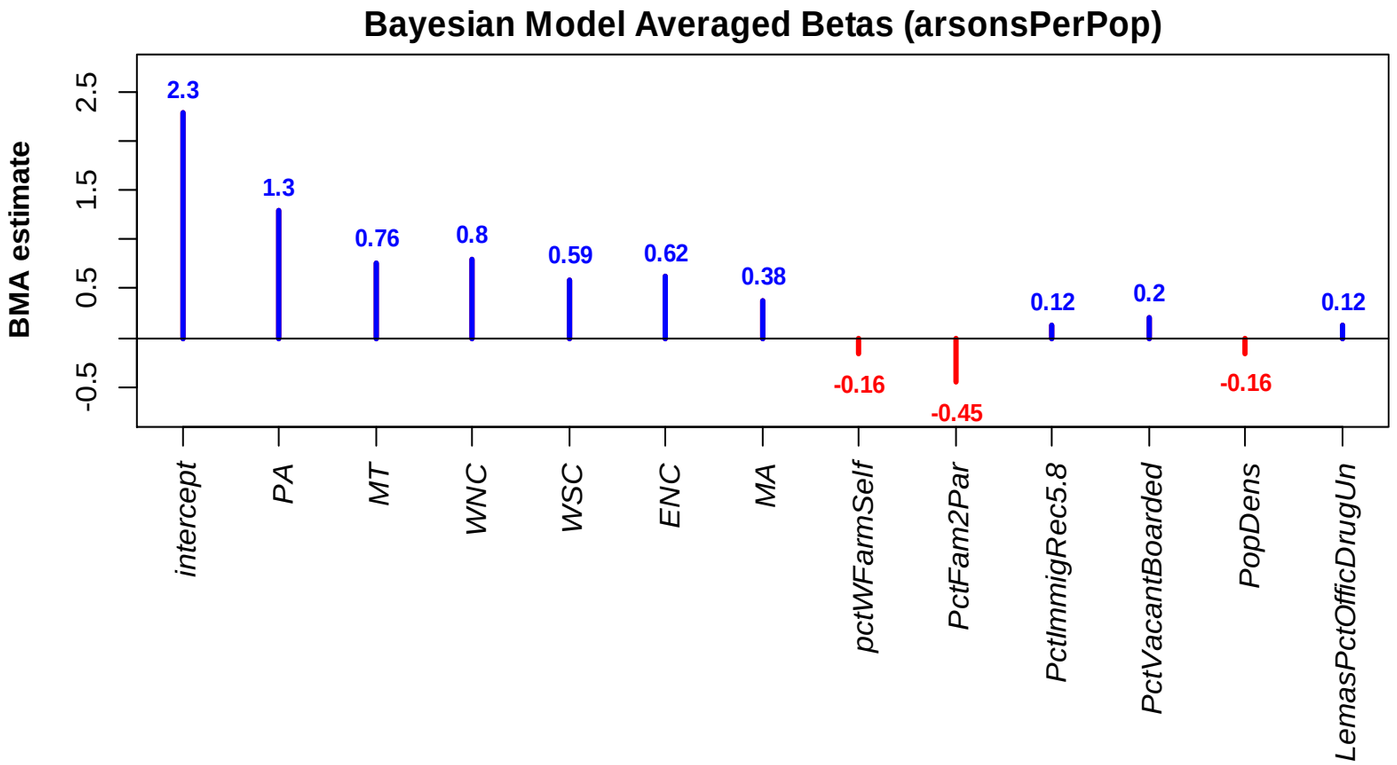


- How to Interpret Gibbs Sampler Result:
  - 깁스 샘플러로 생성한 Z_mean을 보고 predictor를 선정하는 것은 다소 작위적. 우리는 다음과 같은 판단 기준을 세움
  - 가능한 가장 적은 predictor 개수로
    ① Prediction MSE
    ② BIC
    를 최소화하는 Z-cutoff 선택
  - Model Iteration Plot를 보고 Z-cutoff 값과 각 변수의 Z_mean 비교

Model Selection

빈도론적 방법론과 한계

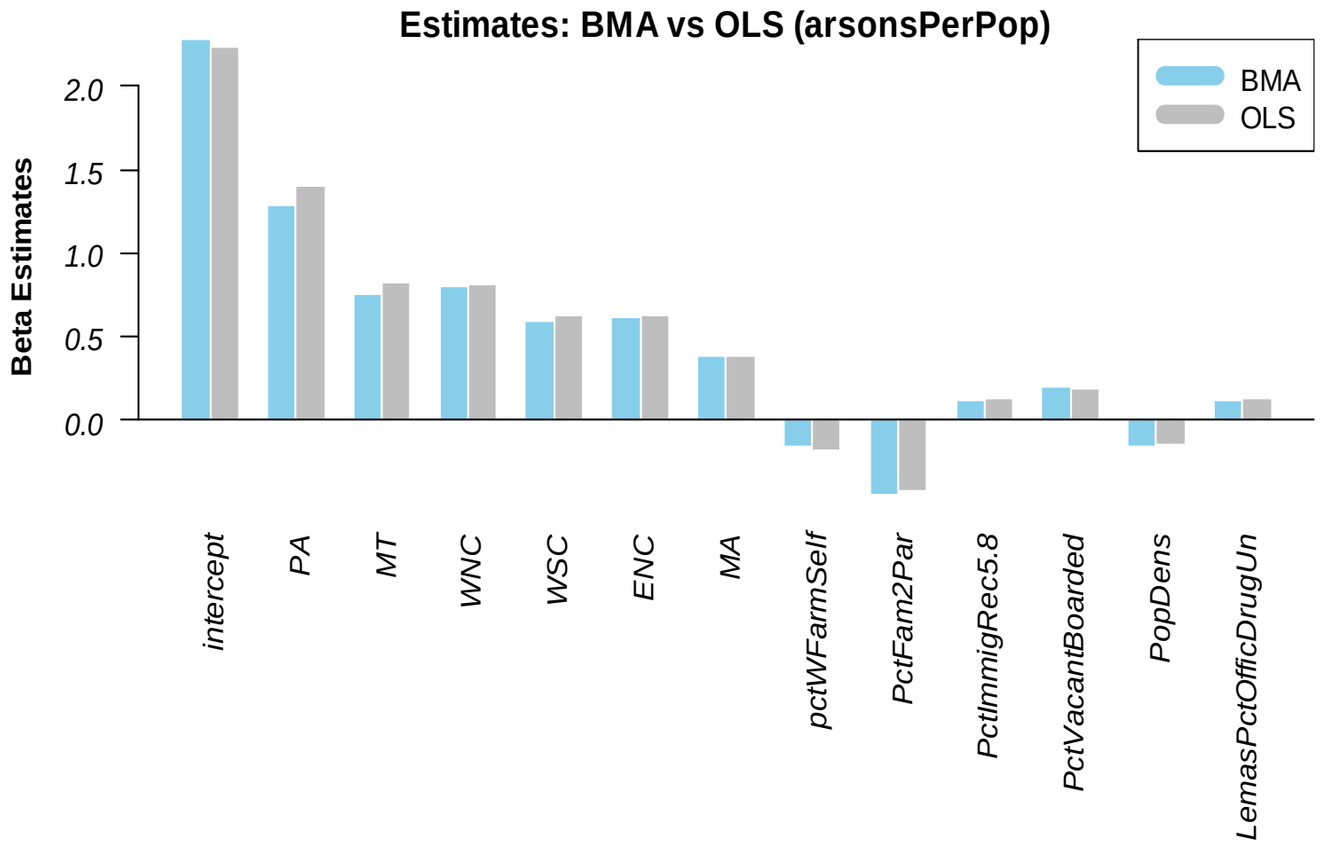
- Bayesian Approach for Model Selection: Model Likelihood



- How to Interpret Gibbs Sampler Result:
  - 깁스 샘플러로 생성한 Z_mean을 보고 predictor를 선정하는 것은 다소 작위적. 우리는 다음과 같은 판단 기준을 세움
  - 가능한 가장 적은 predictor 개수로
  ① Prediction MSE
  ② BIC
  를 최소화하는 Z-cutoff 선택
  - Model Iteration Plot를 보고 Z-cutoff 값과 각 변수의 Z_mean 비교

Model Selection
빈도론적 방법론과 한계

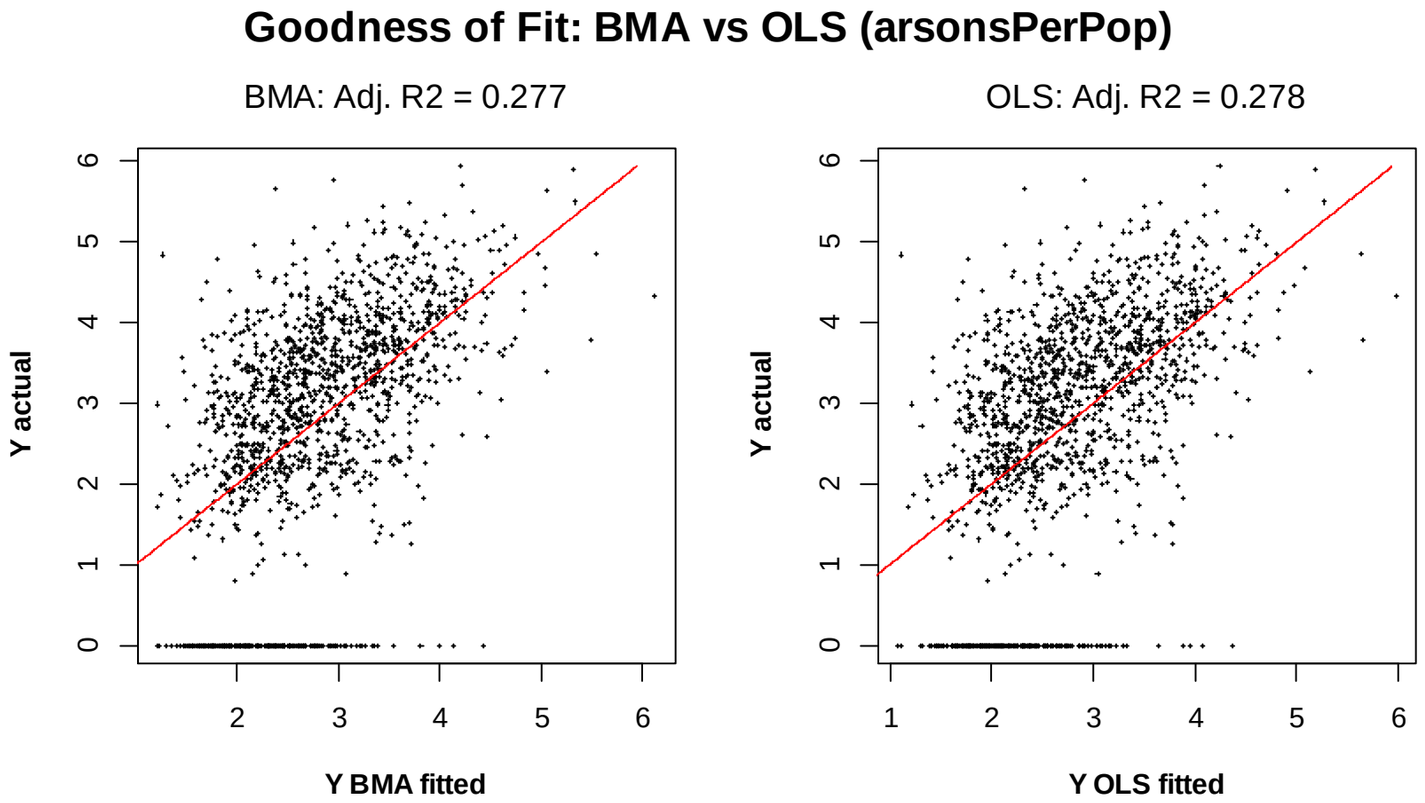
• Bayesian Approach for Model Selection: Model Likelihood



- BMA or OLS Estimator?
  - Bayesian Model Selection으로 적정 변수들을 선택한 후에 이들의 베타를 BMA로 구할지 OLS로 구할지 고민
  - BMA (Bayesian Model Averaging)은, Gibbs Sampler의 결과로 생성된 각각의 모델에 대하여 Beta와 Sigma의 MC approx. 값들의 평균. "가능한 모든 모델에서의 베타의 가중평균"
  - OLS가 주어진 데이터에 Overfitting하는 문제를 완화하는 것으로 알려짐
  - 그러나 BMA와 OLS로 train set fitting, test set prediction 결과에는 큰 차이가 없음 (why? Weak informative prior!)

Model Selection
빈도론적 방법론과 한계

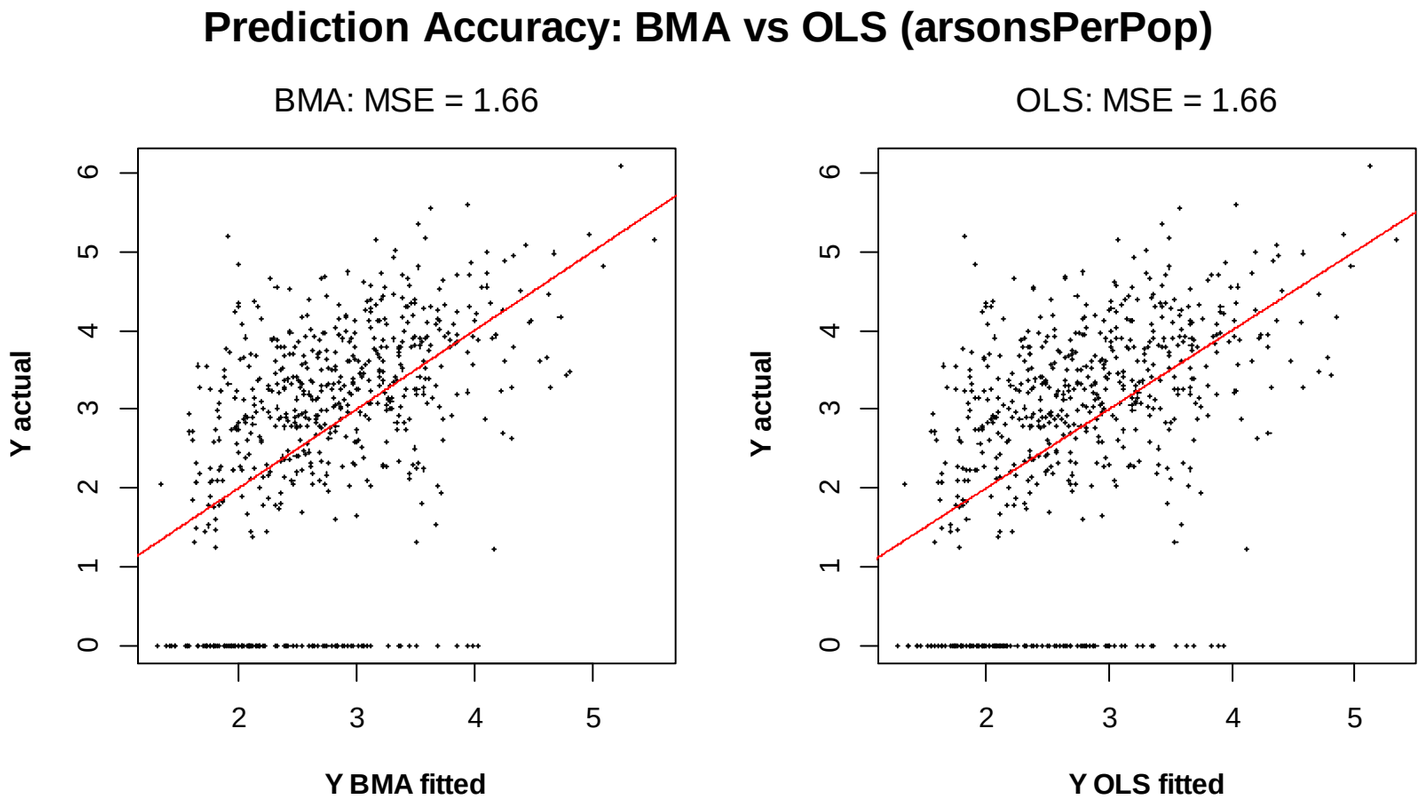
• Bayesian Approach for Model Selection: Model Distribution



- BMA or OLS Estimator?
  - Bayesian Model Selection으로 적정 변수들을 선택한 후에 이들의 베타를 BMA로 구할지 OLS로 구할지 고민
  - BMA(Bayesian Model Averaging)은, Gibbs Sampler의 결과로 생성된 각각의 모델에 대하여 Beta와 Sigma의 MC approx. 값들의 평균. "가능한 모든 모델에서의 베타의 가중평균"
  - OLS가 주어진 데이터에 Overfitting하는 문제를 완화하는 것으로 알려짐
  - 그러나 BMA와 OLS로 train set fitting, test set prediction 결과에는 큰 차이가 없음 (why? Weak informative prior!)

Model Selection
빈도론적 방법론과 한계

- Bayesian Approach for Model Selection: Model Distribution



- BMA or OLS Estimator?
  - Bayesian Model Selection으로 적정 변수들을 선택한 후에 이들의 베타를 BMA로 구할지 OLS로 구할지 고민
  - BMA(Bayesian Model Averaging)은, Gibbs Sampler의 결과로 생성된 각각의 모델에 대하여 Beta와 Sigma의 MC approx. 값들의 평균. "가능한 모든 모델에서의 베타의 가중평균"
  - OLS가 주어진 데이터에 Overfitting하는 문제를 완화하는 것으로 알려짐
  - 그러나 BMA와 OLS로 train set fitting, test set prediction 결과에는 큰 차이가 없음 (why? Weak informative prior!)

V. Model Description : Who did What to Whom?

소득

pctWFarmSelf
pctWPubAsst
whitePerCap
OtherPerCap
PctPopUnderPov
PctEmploy
OwnOccMedVal
MedOwnCostPctIncNoMtg

가구

PctHousNoPhone
PctHousOccup
PctPersOwnOccup
PctLargHouseFam
PctPersDenseHous
PctVacantBoarded
PctVacMore6Mos

이민

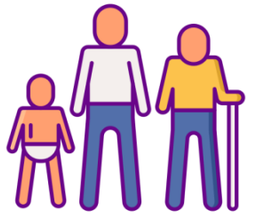
PctForeignBorn
PctImmigRec5.8"
PctSameCity85
PctSpeakEnglOnly

가정환경

TotalPctDiv
PctKidsBornNeverMar
PctWorkMom
PctFam2Par
MalePctNevMarr

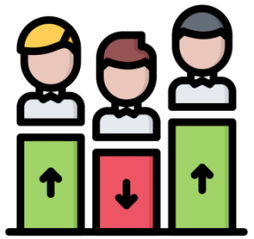
인구

householdsize
PersPerFam
PopDens



연령

agePct12t29
agePct65up



인종

racepctblack
racePctWhite

지역

PA MA ENC MT
WNC WSC NE

기타(도시, 마약, 교육수준)

urban

LemasPctOfficDrugUn

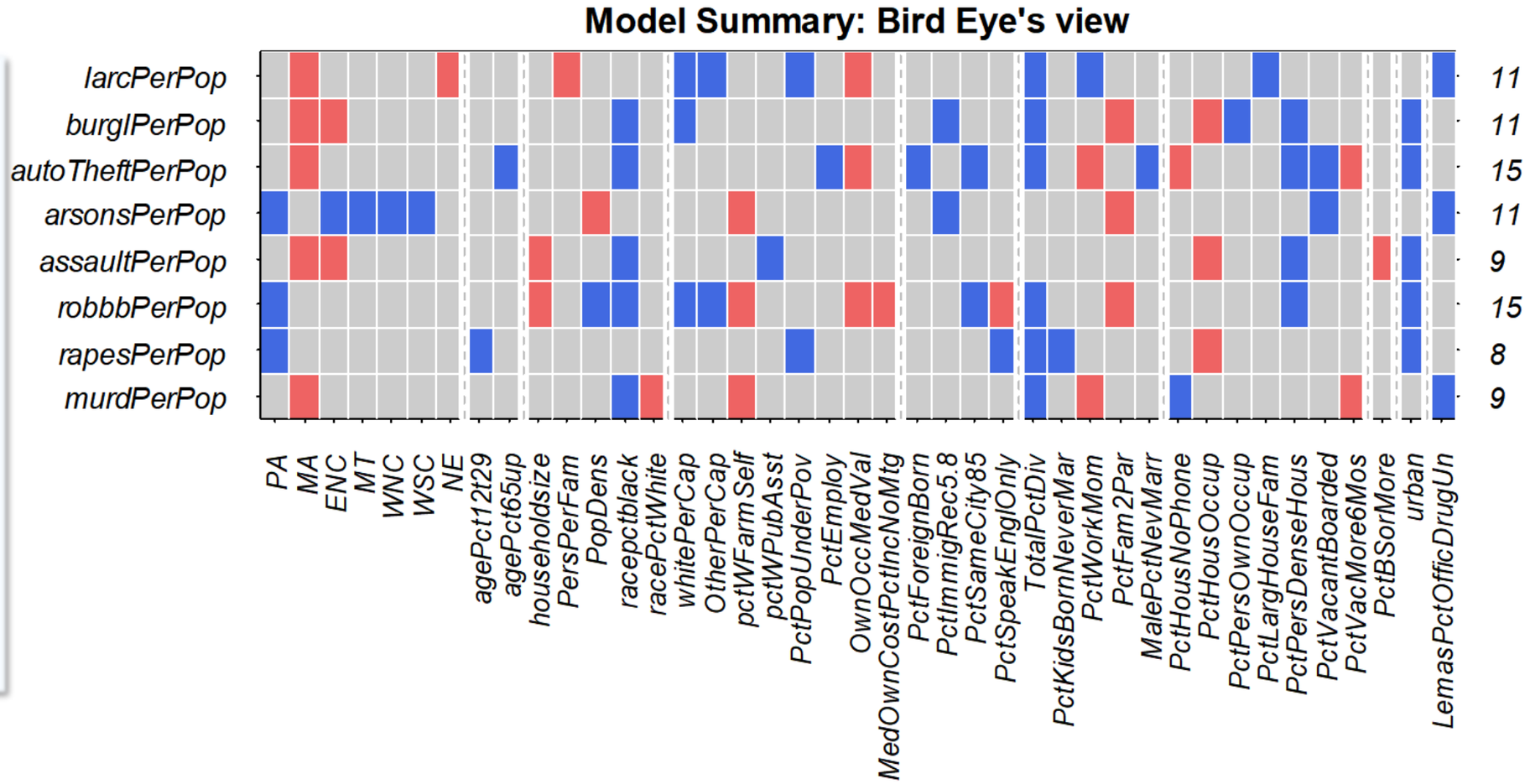
PctBSorMore

Model Selection

빈도론적 방법론과 한계

• Bayesian Approach for Model Selection: Model Distribution

- 총 8개의 모델
- 총 41개의 유의미한 설명변수 (intercept 제외)
- 성격에 따라 분류:
  1. 지역
  2. 인구 밀도
  3. 인종
  4. 소득
  5. 이민
  6. 가정환경
  7. 가구
  8. 기타

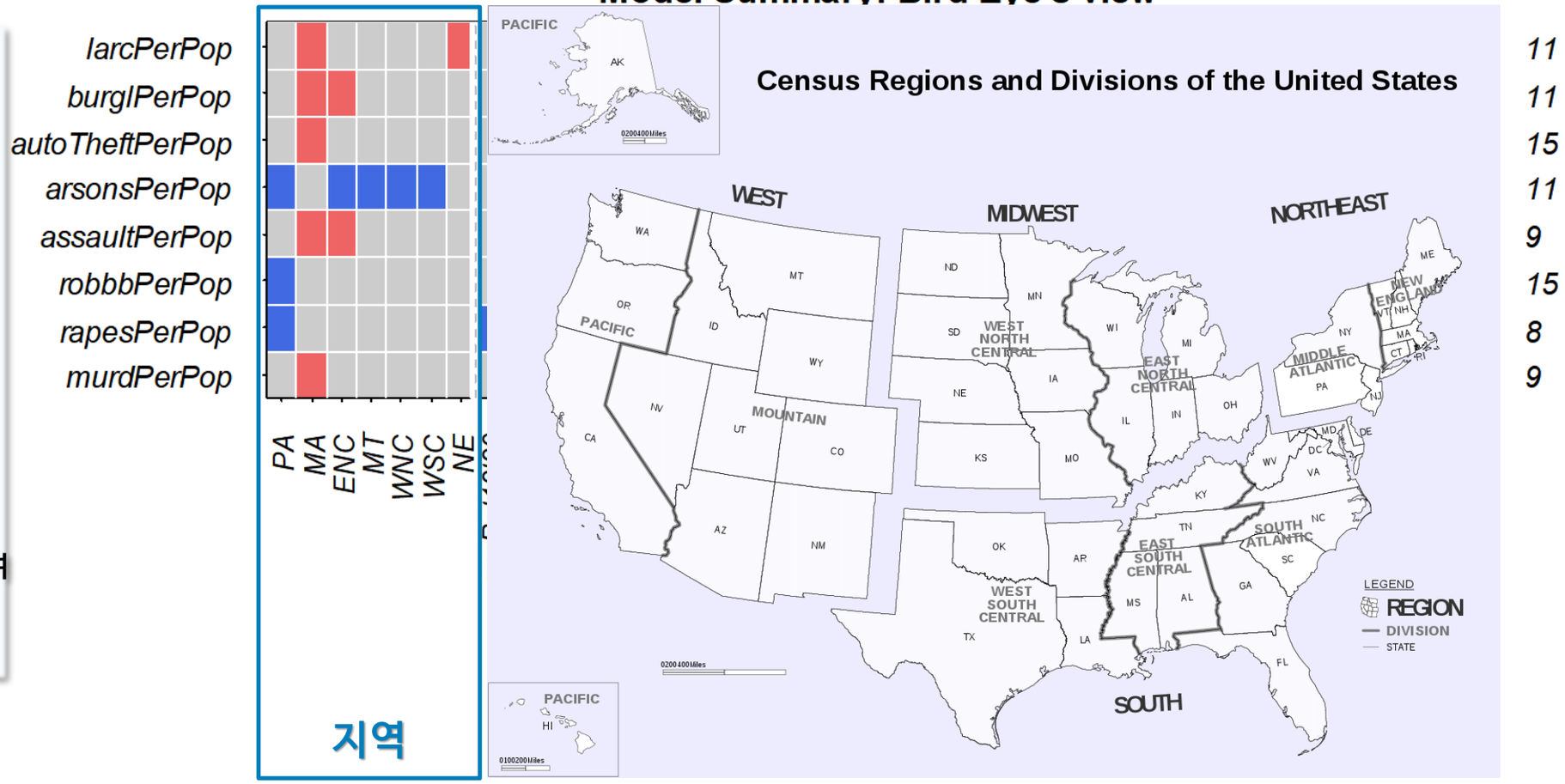


Model Selection

빈도론적 방법론과 한계

• Bayesian Approach for Model Selection: Model Distribution

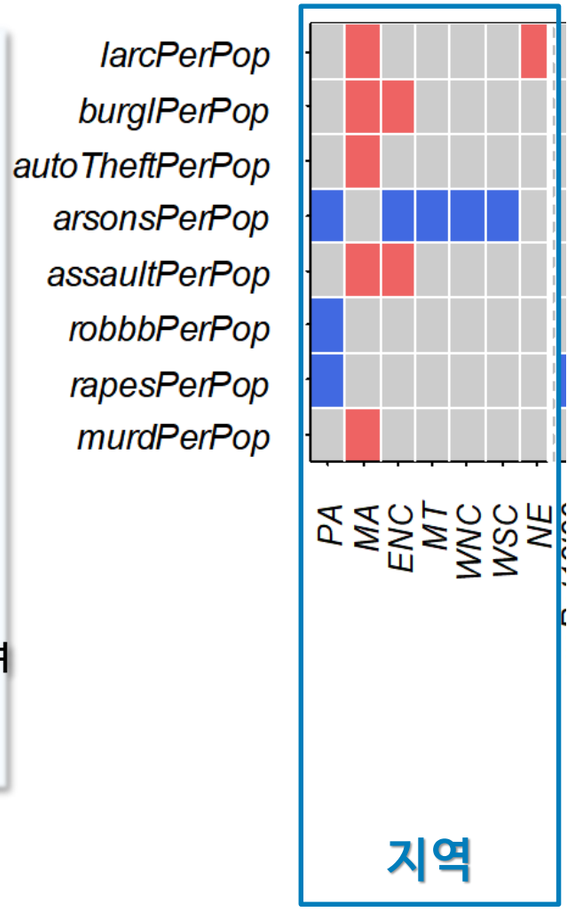
- 지역
  - 각 주를 총 9개의 지역으로 나누어 8개의 더미 변수: (PA, MA, ENC, MT, WNC, WSC, NE, SA)
  - 미 서부 PA는 arsons, robbery, rapes에 대해 *"다른 모든 설명변수의 값이 동일할 때"* 다른 주보다 평균적으로 높음
  - 미 동부 MA는 거의 모든 범죄군에서 평균적으로 범죄율이 낮음
  - 미 중부 MT, WNC, WSC 자체가 arsons 범죄에 대하여 높은 절편을 가짐

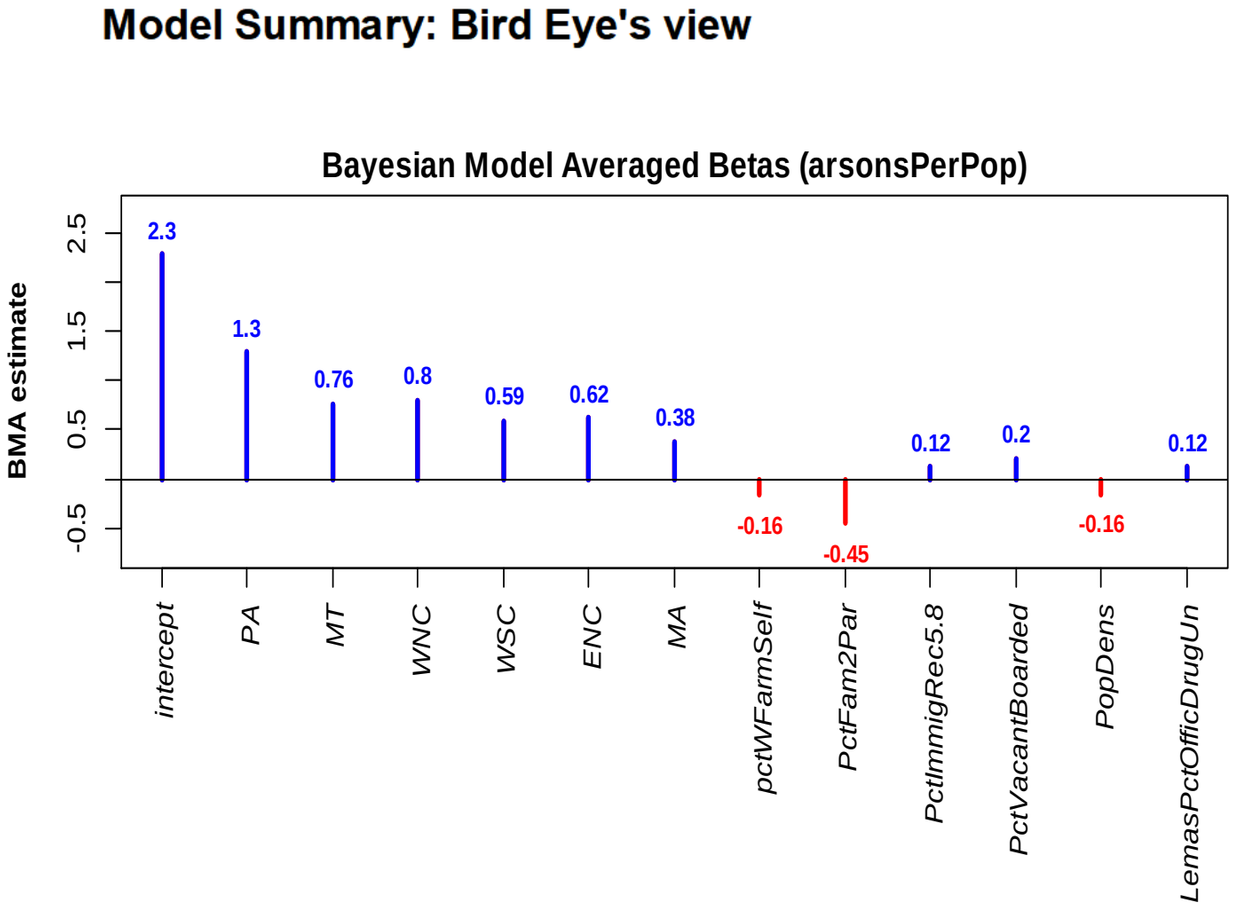


Model Selection
빈도론적 방법론과 한계

• Bayesian Approach for Model Selection: Model Distribution

- 지역
  - 각 주를 총 9개의 지역으로 나누어 8개의 더미 변수: (PA, MA, ENC, MT, WNC, WSC, NE, SA)
  - 미 서부 PA는 arsons, robbery, rapes에 대해 ***"다른 모든 설명변수의 값이 동일할 때"*** 다른 주보다 평균적으로 높음
  - 미 동부 MA는 거의 모든 범죄군에서 평균적으로 범죄율이 낮음
  - 미 중부 MT, WNC, WSC 자체가 arsons 범죄에 대하여 높은 절편을 가짐

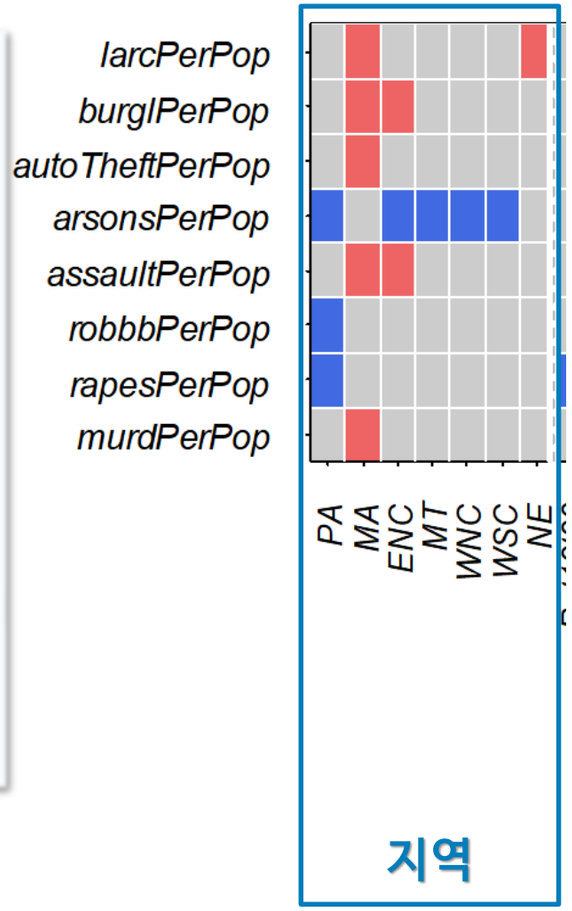




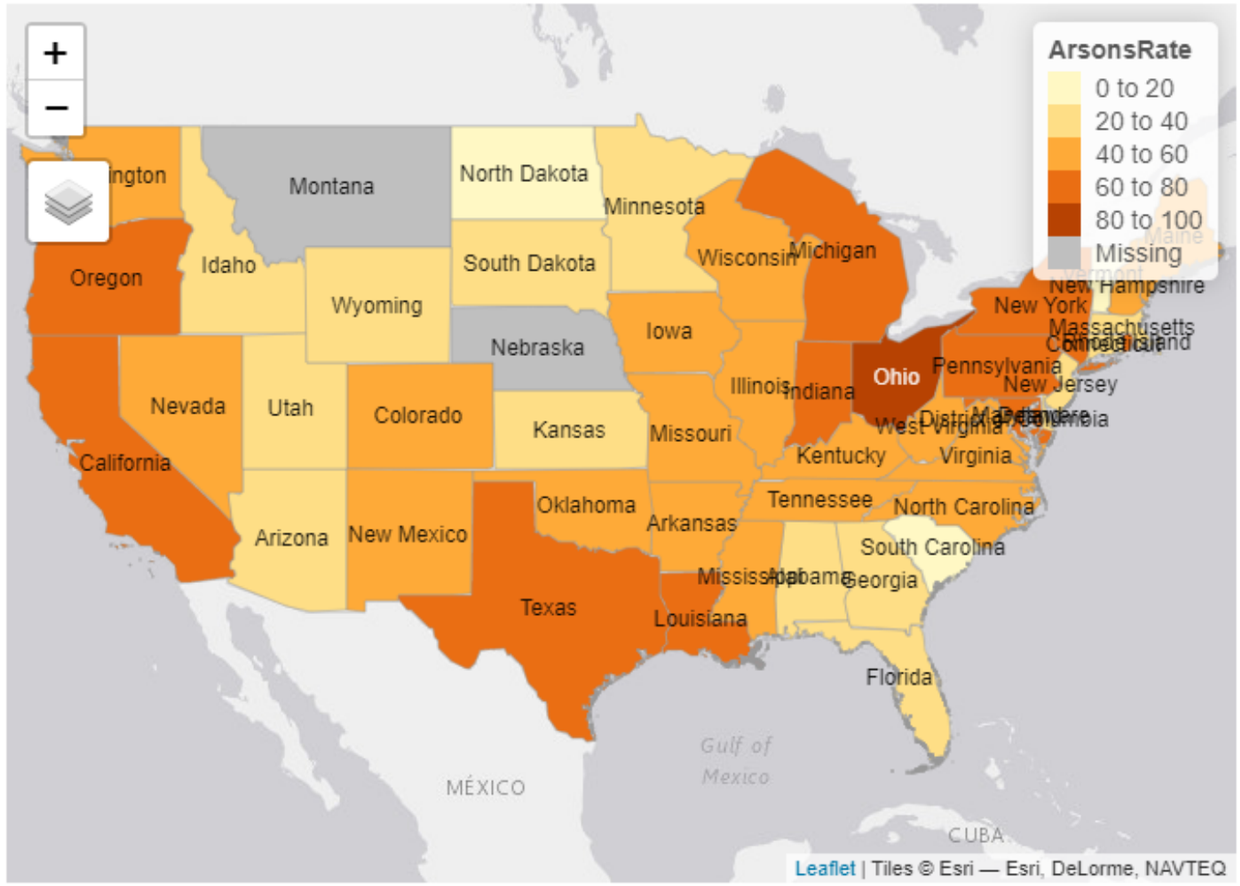
Model Selection
빈도론적 방법론과 한계

• Bayesian Approach for Model Selection: Model Distribution

- 지역
  - 그러나 이것이 MT, WNC, WSC 주의 방화 범죄율이 높다는 것은 아님! 실제로는 PA, MA 지역에서 방화 범죄율이 제일 높음
  - 다만 이는 ***"다른 모든 설명변수의 값이 동일할 때"*** 설명변수의 베타 계수에서 기대되는 정도 보다 높다는 의미.
  - 추측컨대 데이터에 포함되지 않은, 예컨대 산림 비율이나 습도, 강수량, 캠핑장 수 등의 다른 변수가 있을 수 있음.



Model Summary: Bird Eye's view



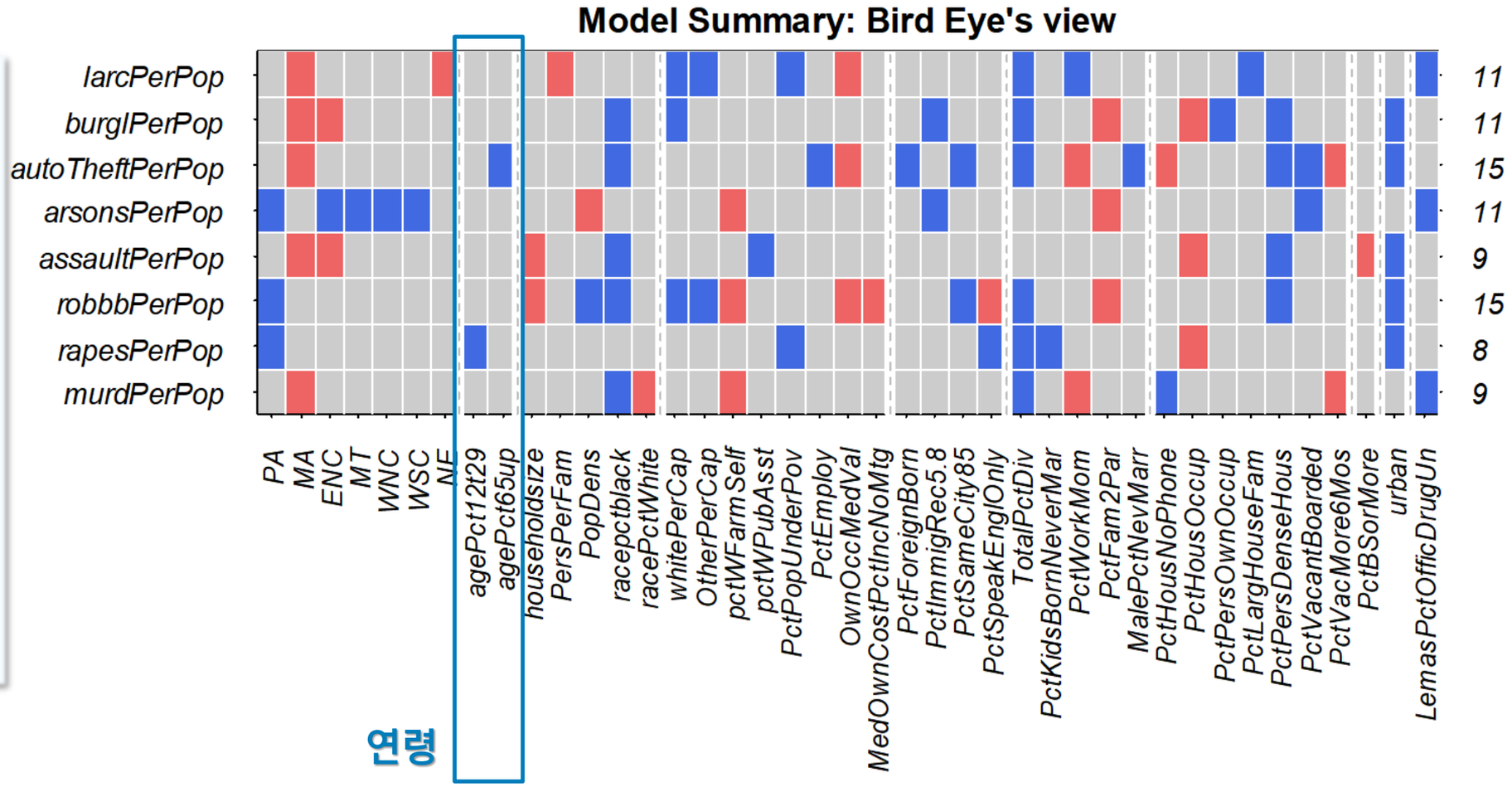
Model Selection

빈도론적 방법론과 한계

• Bayesian Approach for Model Selection: Model Distribution

연령

- 연령 변수는 주로 ***특정 범죄의 대상성***과 관계를 보임
- agePct12t29:
  - 다른 범죄가 아닌 오직 강간에서만 모델에 포함
  - 강간이 주로 특정 연령층을 대상으로 이뤄지고 있다..!
- agePct65up:
  - 차량 강탈 모델에만 포함
  - 아무래도 신체적으로 약자인 노년층이 차량 강탈의 주된 타겟이 되는 것이 아닌가!



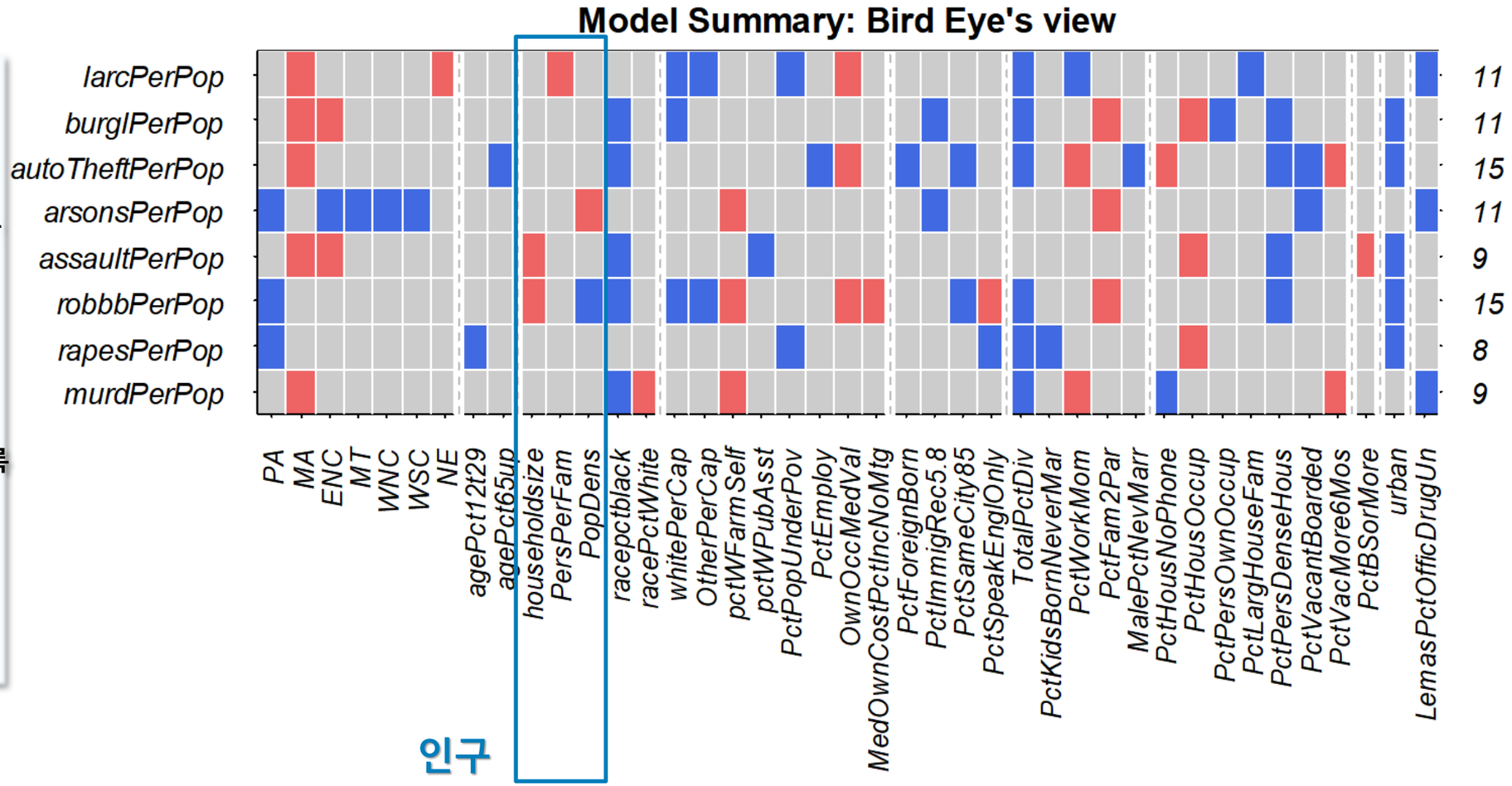
Model Selection

빈도론적 방법론과 한계

- Bayesian Approach for Model Selection: Model Distribution

인구

- 인구 밀집도는 larceny, assault, arson 등의 범죄율을 감소하는 것으로 나타남
  - 인구 밀집도가 높을수록 치안 및 소방 인프라가 잘 갖춰져 있다?
- 그러나 robbery의 경우 PopDens가 범죄율을 높이는 것으로 나타남.
  - 이는 인구 밀집도가 높을 수록 강도의 대상이 되는 각종 상점이 많아서 그럴 것이라 추측



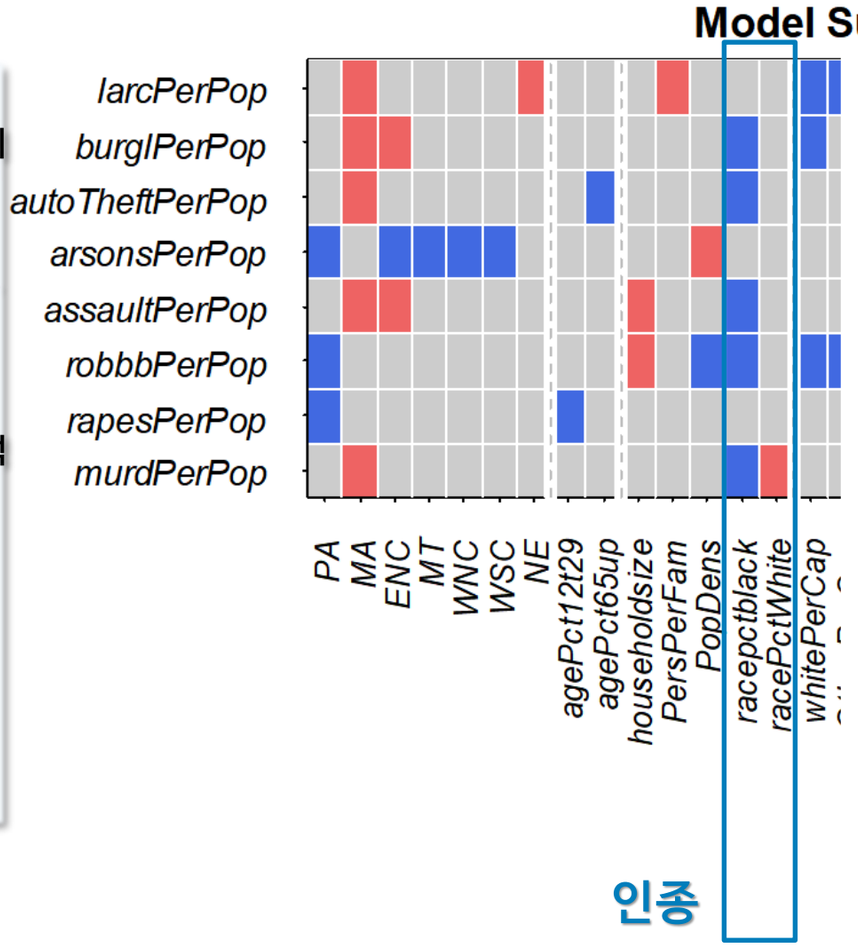
Model Selection

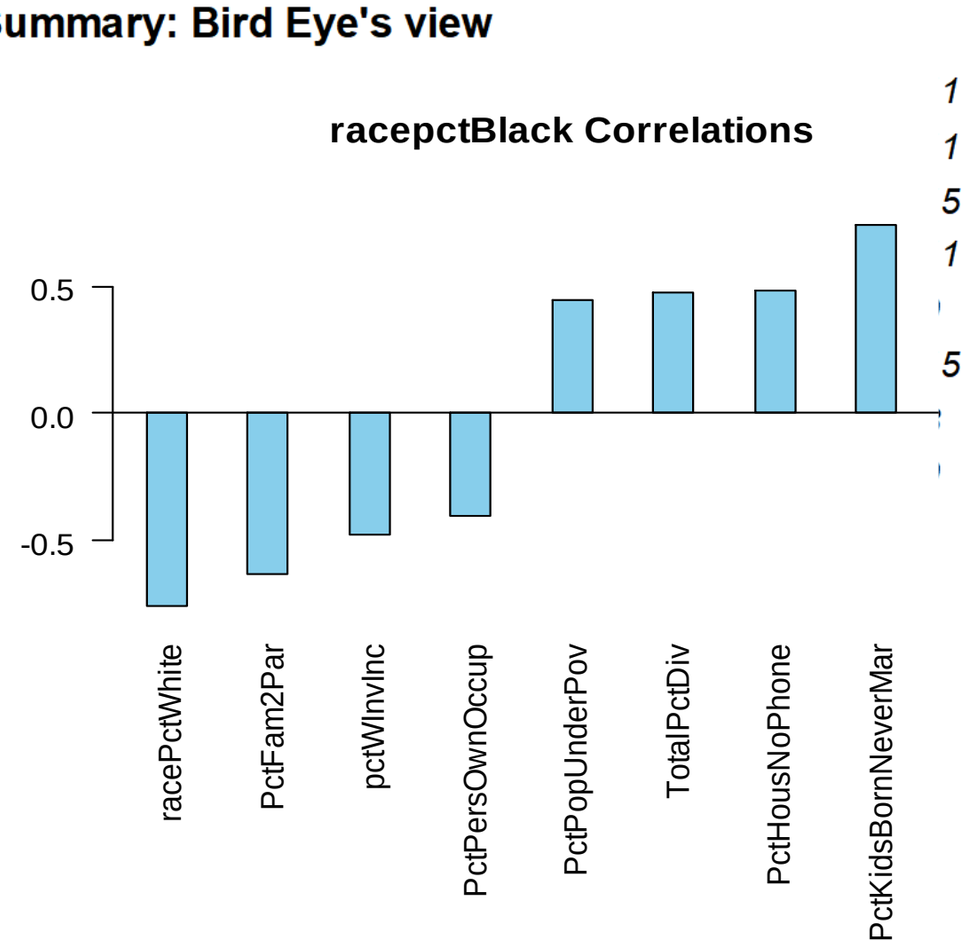
빈도론적 방법론과 한계

• Bayesian Approach for Model Selection: Model Distribution

인종

- 흑인 비율은 거의 모든 범죄군의 범죄율에 기여 *"흑인 나빠요?"*
  - No. 흑인 거주지역의 사회경제적 인프라가 열악함. 특히 불우한 가정환경을 나타내는 변수와 높은 상관성을 가짐
  - 미국의 뿌리 깊은 인종차별적 정책의 영향을 고려해야..
- 백인 비율은 오직 살인에서만 음의 기여를 함
  - 가장 심각한 강력범죄가 살인임을 감안하면 백인 거주지역의 치안 수준이 높다고 해석할 수도..!





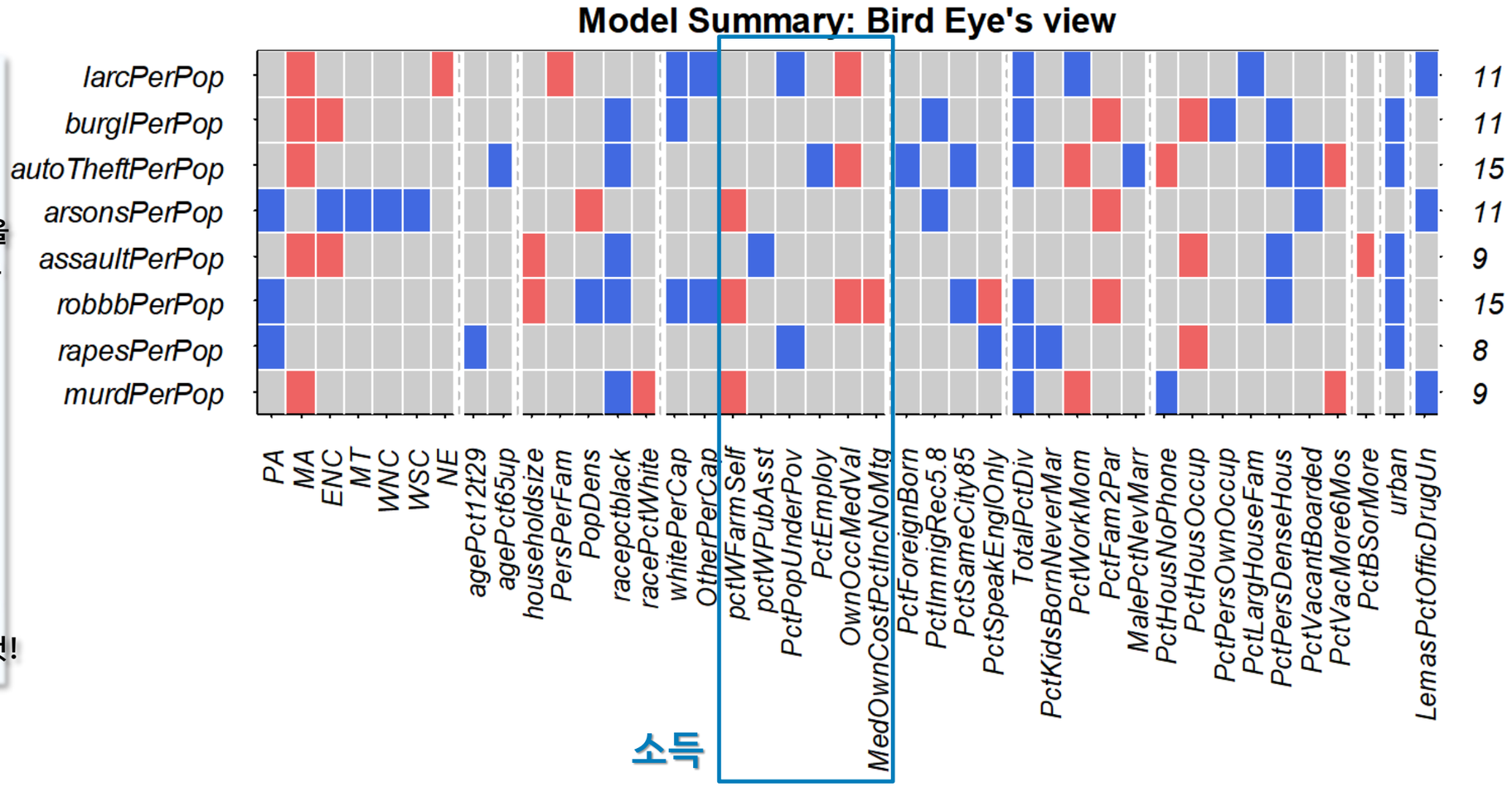
Model Selection

빈도론적 방법론과 한계

• Bayesian Approach for Model Selection: Model Distribution

소득

- 소득 변수의 범죄율 영향은 대체로 직관과 일치
  - 자가소득자 비율, 미디언 주거비용 등 높은 소득 수준을 짐작하는 변수는 범죄율 감소
  - 기초수급자, 빈곤선 이하 비율 등의 변수는 범죄율 증가에 기여
- 다만 고용률과 차량절도는 어리둥절. 하지만 agePct65up을 생각한다면 "훔칠 차가 많아서?"
  - Okun's Law를 생각하면 고용률은 지역 경기의 Proxy
  - 그만큼 (훔칠) 차량이 많을 것!



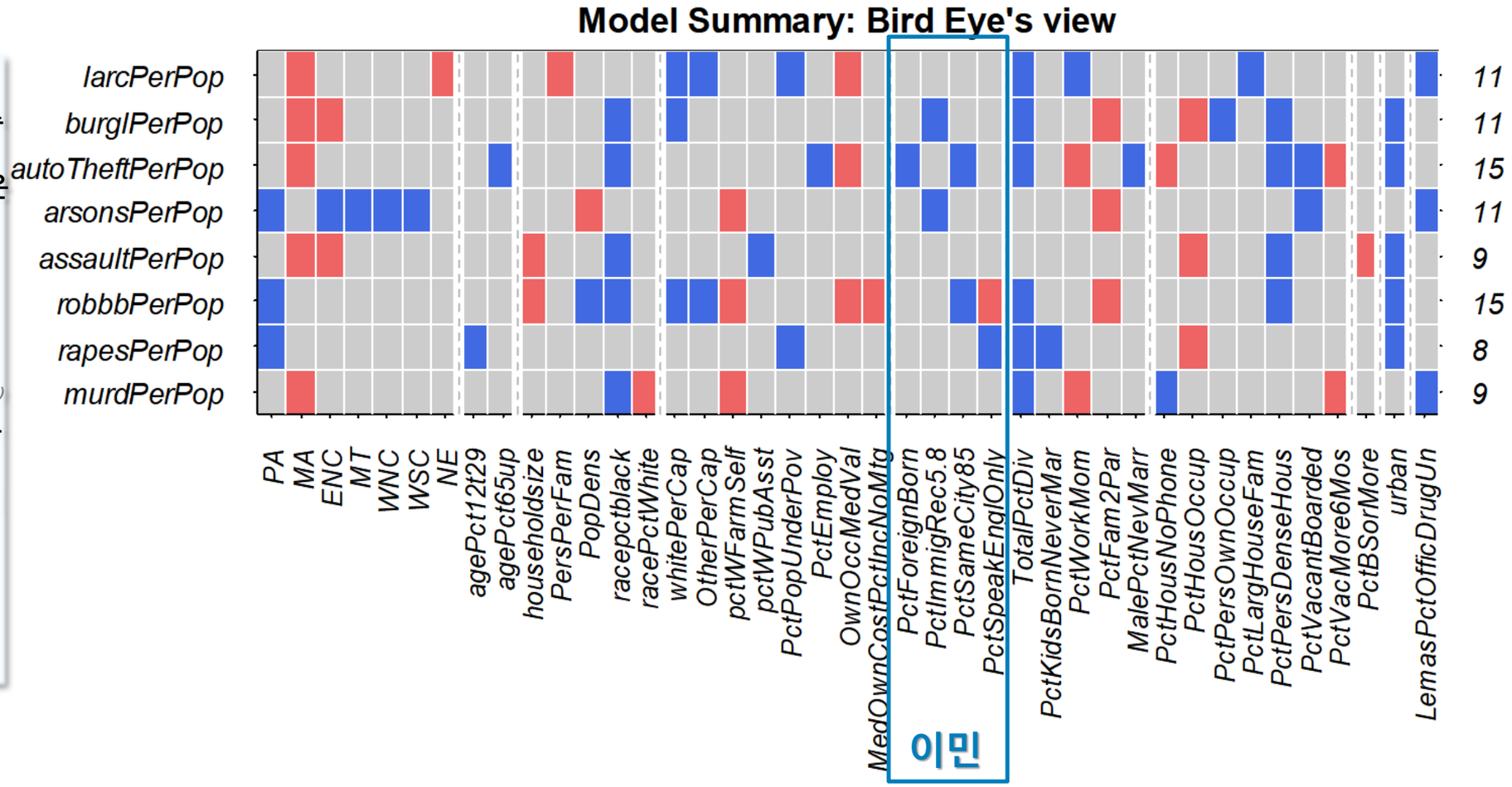
Model Selection

빈도론적 방법론과 한계

• Bayesian Approach for Model Selection: Model Distribution

이민

- 외국 태생 비율, 85년 이후 거주 비율, 5~8년 전 이민자 비율은 이민자, 영어만 가능자의 비율은 원주민에 관련된 변수로 간주
- 이민자 비율은 주로 생계형 범죄(절도, 차량강탈, 강도)
  - 주로 히스패닉 계열
    (70년대 이후 미국의 마약 사용이 급증하면서 카르텔 부상에 따른 사회 불안으로 대거 미국 이주)
  - 이들이 사회 적응에 어려움을 겪음을 짐작할 수 있음. 5~8년 비율은 "적응 실패"의 proxy로 볼 수 있다..?



Model Selection

빈도론적 방법론과 한계

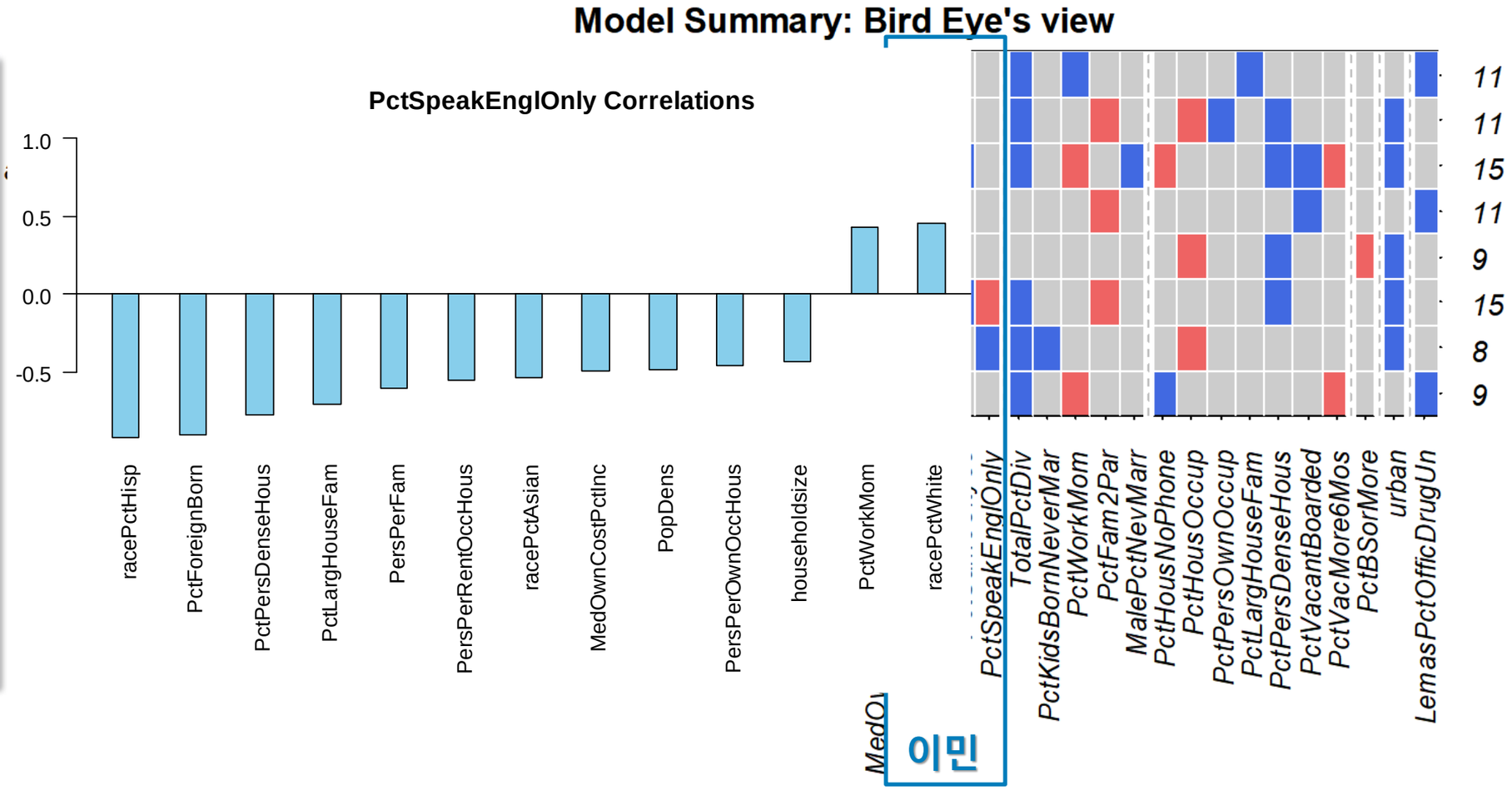
• Bayesian Approach for Model Selection: Model Distribution

이민

- **영어만 쓰면 넌 강간마..?**
  상관관계를 고려하면 주로
  - **① 미국 태생의 *백인***
  - **② 가구 구성원 적음**
  - **③ (다소) 고소득**
  - **④ *맞벌이* 가정**
- **가정 교육을 제대로 못 받은**
  백인 중산층 자녀..?

  상대적으로 소득이 안정되지만
  가정에서 보내는 시간이 적은
  백인 중산층 부모..?

(어디까지나 추측일 뿐!)



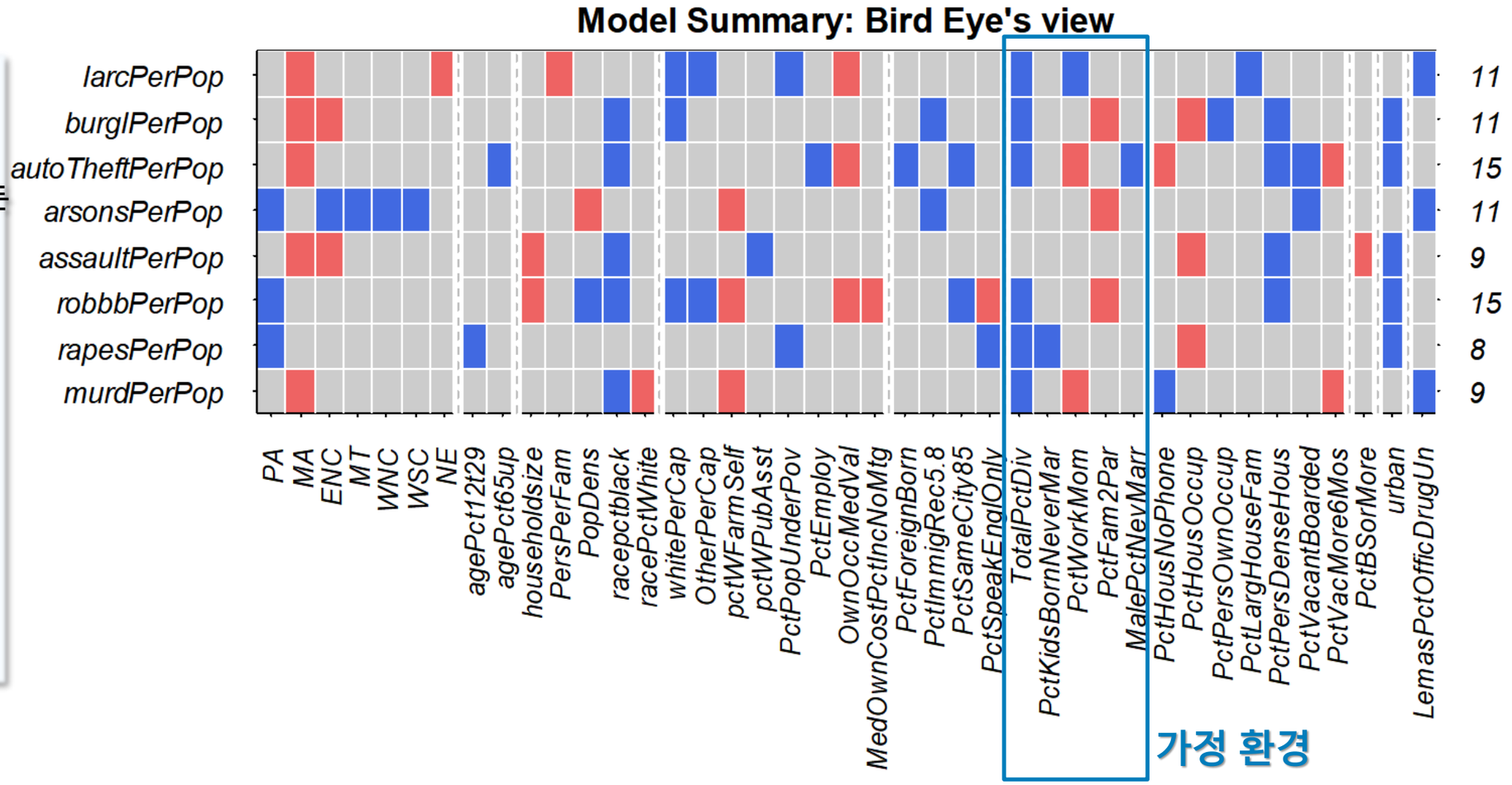
Model Selection

빈도론적 방법론과 한계

• Bayesian Approach for Model Selection: Model Distribution

가정 환경

- 가정 환경의 범죄율 영향은 대체로 직관과 일치
  - 특히 이혼 가정의 비율은 모든 범죄율에 기여, 양부모 가정 범죄율 감소 영향
  - PctWorkMom: 아이들이 소매치기를..?
- MalePctNevMarr: 주로
  - ① 최근에 이민을 온
  - ② 21~29세의 젊은
  - ③ 소득 수준이 낮은
  - ④ 미혼모 자녀
  - 차량절도의 주범으로 짐작



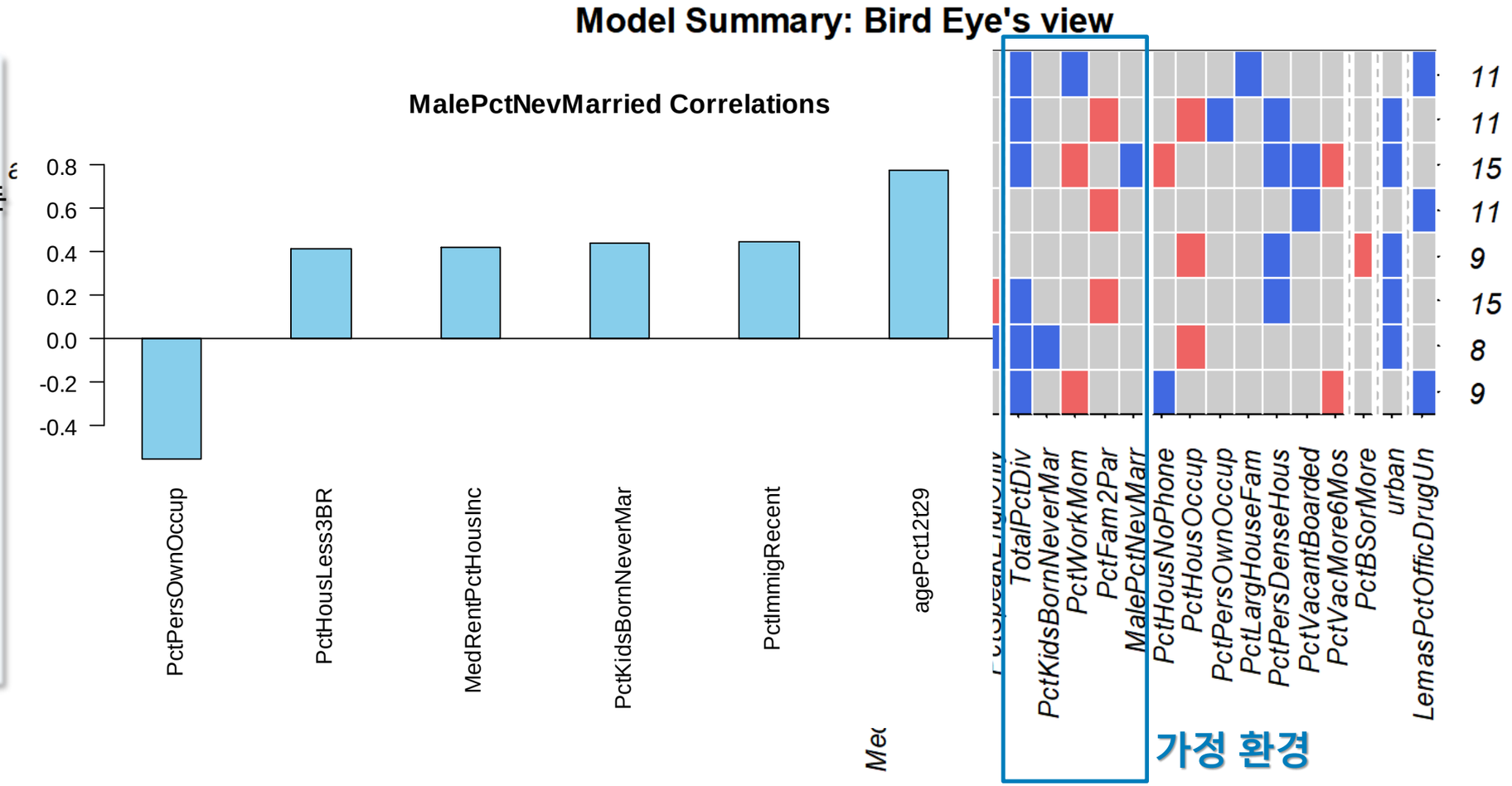
Model Selection

빈도론적 방법론과 한계

• Bayesian Approach for Model Selection: Model Distribution

가정 환경

- 가정 환경의 범죄율 영향은 대체로 직관과 일치
  - 특히 이혼 가정의 비율은 모든 범죄율에 기여, 양부모 가정 범죄율 감소 영향
  - PctWorkMom: 아이들이 소매치기를..?
- MalePctNevMarr: 주로
  - ① 최근에 이민을 온
  - ② 21~29세의 젊은
  - ③ 소득 수준이 낮은
  - ④ 미혼모 자녀
  - 차량절도의 주범으로 짐작



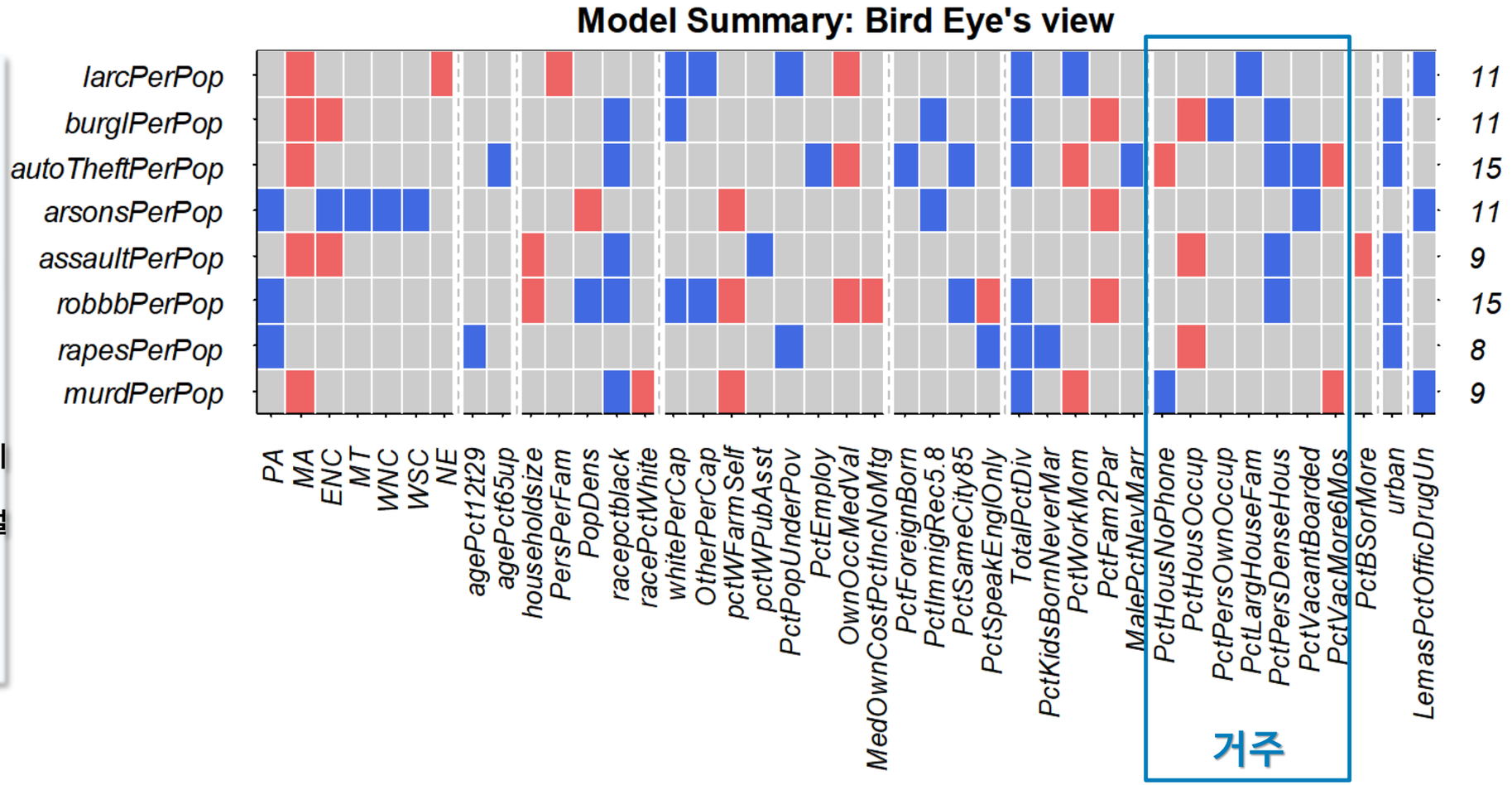
Model Selection

빈도론적 방법론과 한계

• Bayesian Approach for Model Selection: Model Distribution

거주 (= 소득 수준)

- PctHousOccup:
  빈집 비율의 반대로, 전반적인 지역 경기를 나타낸다고 간주
- PctPersDensHous:
  Dense Housing은 우리나라의 아파트와 비슷하지만, 미국은 주로 Project라 불리며 빈민층이 많이 거주함
- PctVacantBoarded:
  아무도 안 사는 빈집을 집주인이 판자로 보강을 해 놓은 비율. 지역의 낮은 치안 수준을 나타낼 수 있음



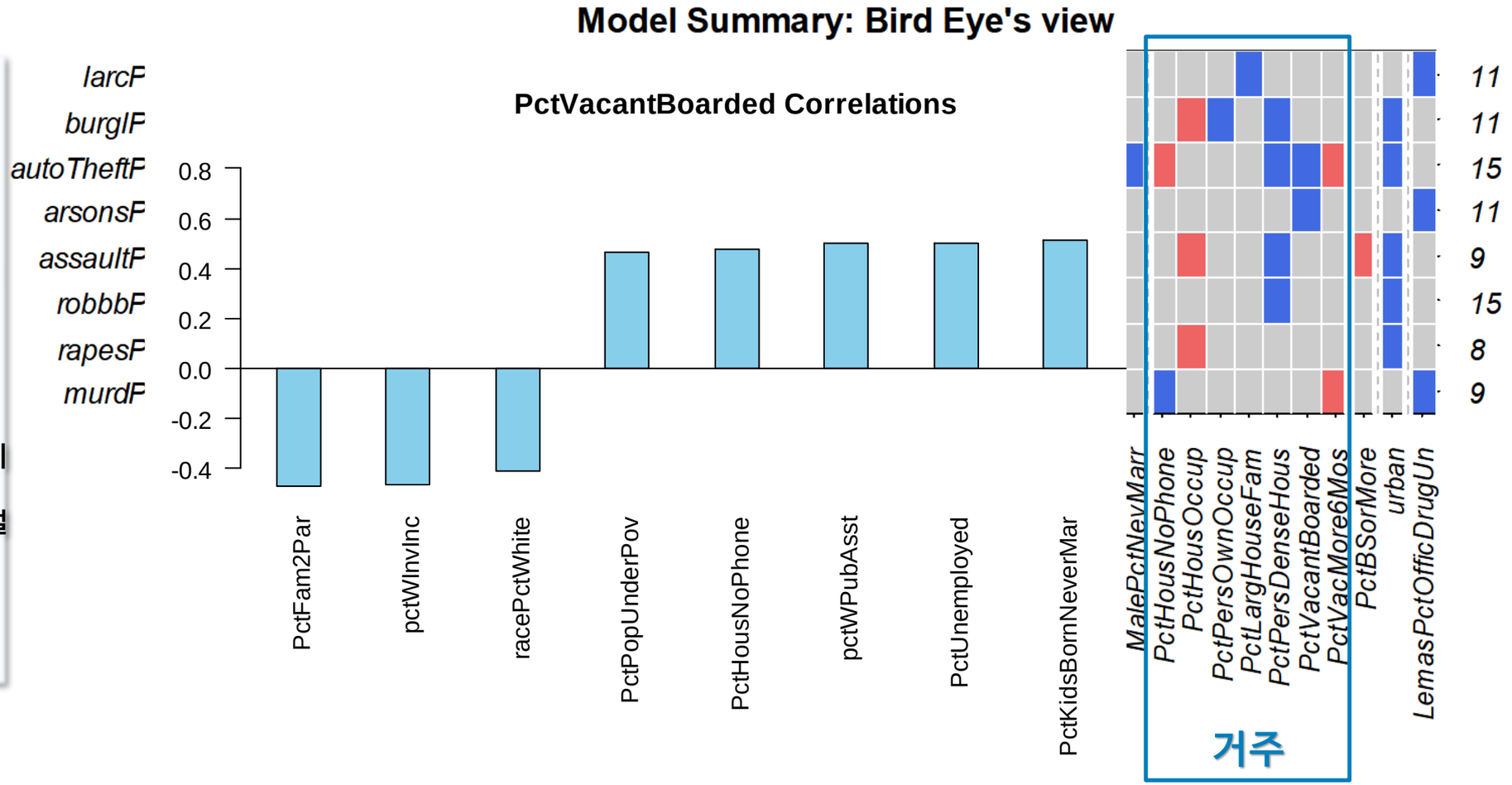
Model Selection

빈도론적 방법론과 한계

• Bayesian Approach for Model Selection: Model Distribution

거주 (= 소득 수준)

- PctHousOccup:
  빈집 비율의 반대로, 전반적인 지역 경기를 나타낸다고 간주
- PctPersDensHous:
  Dense Housing은 우리나라의 아파트와 비슷하지만, 미국은 주로 Project라 불리며 빈민층이 많이 거주함
- PctVacantBoarded:
  아무도 안 사는 빈집을 집주인이 판자로 보강을 해 놓은 비율. 지역의 낮은 치안 수준을 나타낼 수 있음



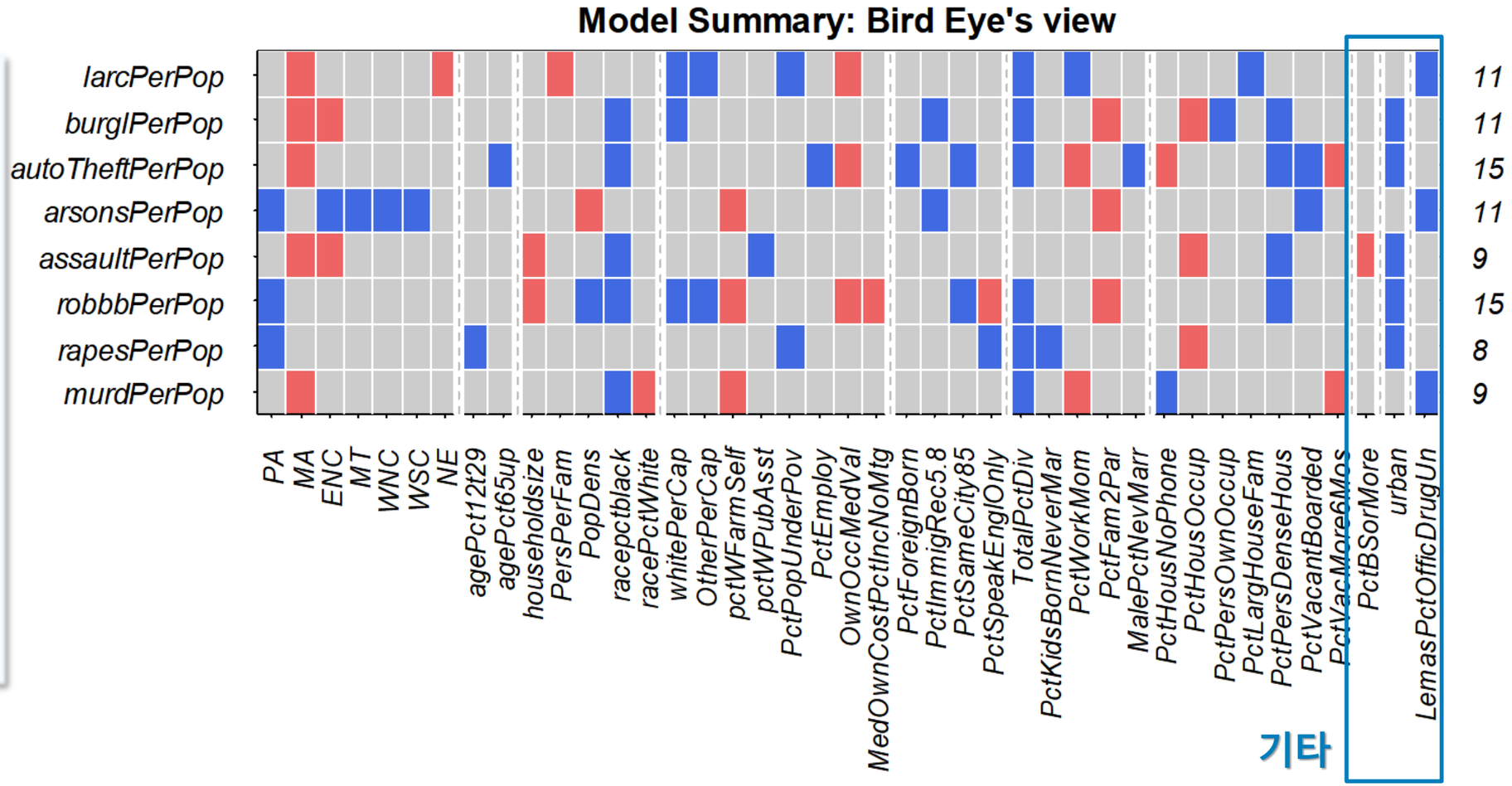
Model Selection

빈도론적 방법론과 한계

- Bayesian Approach for Model Selection: Model Distribution

기타

- urban:
  도시 거주 인구 비율을 0, 1로 binary 변환. 도시가 대부분의 범죄율이 더 높다.



V. Graphical Representation (WARNING: GRAPHIC CONTENT)

Q&A
Thank You