『Marketing Science』

Week 02 : Bayesian Rethinking

권남택 | Netmarble

목차

Causal-Lab

Are you Bayesian?

Causal-Lab

베이지안을 정의하자면…

- 빈도주의 (Frequentist) / 베이지안 (Bayesian)
- 빈도주의 : 특정 사건에 대한 무한한 반복을 가정함으로써 추론을 시도
  - ex. $X_1, \dots, X_n \sim NID(0,1)$, 확률표본 (random sample), 상대도수의 극한, …
  - 모수 θ (parameter) : Fixed but unknown constant
  - 데이터의 점근적 특징 (Asymptotic Property) 에 관심. $n \geq 30$ 이라면 표본평균은 정규분포…
- 베이지안 : 주어진 데이터를 기반으로 불확실성을 고려함으로써 추론을 시도
  - ex. 주어진 데이터만 가지고, 사전분포와 결합함으로써 모수에 대한 추론
  - 모수는 고정된 값이 아니라, 불확실성을 지니는 확률변수

Are you Bayesian?

Causal-Lab

빈도주의 추론의 맹점

- 최대 가능도 추정 – 빈도주의의 대표적인 추정 방법
- 동전던지기에서 앞면이 나올 확률 p 에 대한 최대가능도추정량은 표본평균 $\bar{X}$
- 만약 동전을 한 번 던져서 앞면이 1번 나왔을 때,

  p에 대한 추정치는?

참고자료 : builtin

Are you Bayesian?

Causal-Lab

빈도주의 추론의 맹점

- 최대 가능도 추정 – 빈도주의의 대표적인 추정 방법
- 동전던지기에서 앞면이 나올 확률 p 에 대한 최대가능도추정량은 표본평균 $\bar{X}$
- 만약 동전을 한 번 던져서 앞면이 1번 나왔을 때,

  p에 대한 추정치는?

앞면이 나올 확률이 1이라고 할 수 있을까?

참고자료 : builtin

Are you Bayesian?

Causal-Lab

빈도주의 추론의 맹점

- 동전을 한 번 던져서 앞면이 1번 나왔을 때, 확률을 1이라고 할 수도 있지만…
- 1이 아닌 다른 어떤 값을 생각한다면…

  당신은 Bayesian적 사고를 이미 하고 있습니다
- 베이지안은 단순한 데이터를 넘어, 보다 합리적인 추론을 가능하게 함
- 이론적으로는 불확실성에 대한 논의와 관련되어 있음
  - 우연적 불확실성 vs 인식적 불확실성

참고자료 : 오하이오의 낚시꾼 게시글 - [두 종류의 불확실성: 우연적 vs 인식적 불확실성]

Bayesian is all you need?

Causal-Lab

베이지안이어야 하는가?

- 우리 사고체계 자체는 베이지안에 더 가깝다는건 늘 주장됨
- 빈도주의적 관점의 iid 무한 반복은 많은 도메인에서 허용되기 어려움
  - 빈도주의적 추론의 신뢰구간 / p-value.. 우리가 알고싶은 대상인가?
  - 시간에 따라 경쟁/효과가 변하는데, 반복을 가정할 수 있는가?

Bayesian is not all you need

Causal-Lab

All Model are wrong, but some are useful.

- George Box (1976)
- 베이지안 관점이 옳다기보다, 베이지안 모델링이 가져올 수 있는 장점에 집중
  - 불확실성에 대한 계량
  - 구조적 특징에 대한 부여
  - 온라인 업데이트 가능 + 결과에 대한 해석

Link with Causal Inference

Causal-Lab

베이지안과 인과추론

- 베이지안과 인과추론은 데이터를 바라보는 관점 (Framework)으로, 충돌되지 않음
- 빈도주의 + 인과추론은 보통 잠재적 결과 (Potential Outcome; PO)에서만 이해하지만,

  베이지안 + 인과추론은 PO + Graph로 문제를 정의하는 경향이 있음 (베이지안 네트워크)
- 미국에는 베이지안 기반 인과추론 연구가 매우 활발

  - Richard Hahn : 베이지안 트리앙상블 기반 인과추론 (Bayesian Causal Forest, …)

  - 베이지안 트리앙상블은 인과추론계의 SOTA

Link with Causal Inference

Causal-Lab

인과추론을 위한 베이지안

- 불확실성 계량 : 사람과 관련된 연구주제가 많아, 효과 추정을 넘어 효과에 대한 불확실성 계량 니즈

  - 빈도주의도 가능하긴 하지만, 이를 통합적 / 해석 측면에서 합리적 모델링

- 편향을 유발하지 않는 추론 구조 부여

  - MMM : Ridge는 편향을 발생시킨다는 비판 => Bayesian 선형회귀로 해결

  - 머신러닝스러운 모델링에서도 상대적으로 편향을 보정하는 구조들을 가져가기 더 쉬움

Basic Bayesian Terms

베이지안 추론의 형태

- 사후분포 (posterior) $\propto$ 가능도 (likelihood, 데이터) * 사전분포 (prior)
- 데이터는 주어진 값, 관심있는 확률변수는 θ. 베이즈 정리를 통해 전개
- 가능도와 사전분포가 특수한 경우, 매우 간단한 꼴로 정리된다는게 알려져 있음
  - ex. 평균에 대한 추론 : 정규분포 (가능도) * 정규분포 (사전분포) => 정규분포 (사후분포)
  - ex. 확률에 대한 추론 : 이항분포 (가능도) * 베타분포 (사전분포) => 베타분포 (사후분포)

$$f(\theta|x) = \frac{p(\theta,x)}{p(x)} = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta} = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta} \propto p(x|\theta)p(\theta)$$

Basic Bayesian Terms

Causal-Lab

When likelihood function is a continuous distribution [edit]

Small
Standard

Likelihood $p(x_i\|\theta)$	Model parameters θ	Conjugate prior (and posterior) distribution $p(\theta\|\Theta), p(\theta\|\mathbf{x},\Theta) = p(\theta\|\Theta')$	Prior hyperparameters Θ	Posterior hyperparameters[note 1] Θ'	Interpretation of hyperparameters	Posterior predictive[note 5] $p(\tilde{x}\|\mathbf{x},\Theta) = p(\tilde{x}\|\Theta')$
Normal with known variance σ^2	μ (mean)	Normal	$\mu_0,\ \sigma_0^2$	$\frac{1}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_{i=1}^n x_i}{\sigma^2}\right), \left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}$	mean was estimated from observations with total precision (sum of all individual precisions) $1/\sigma_0^2$ and with sample mean μ_0	$\mathcal{N}(\tilde{x}\|\mu_0', \sigma_0^{2\prime}+\sigma^2)$[4]
Normal with known precision τ	μ (mean)	Normal	$\mu_0,\ \tau_0^{-1}$	$\frac{\tau_0\mu_0+\tau\sum_{i=1}^n x_i}{\tau_0+n\tau}, (\tau_0+n\tau)^{-1}$	mean was estimated from observations with total precision (sum of all individual precisions) τ_0 and with sample mean μ_0	$\mathcal{N}\left(\tilde{x}\mid\mu_0', \frac{1}{\tau_0'}+\frac{1}{\tau}\right)$[4]
Normal with known mean μ	σ^2 (variance)	Inverse gamma	$\alpha,\ \beta$ [note 6]	$\alpha+\frac{n}{2},\ \beta+\frac{\sum_{i=1}^n (x_i-\mu)^2}{2}$	variance was estimated from 2α observations with sample variance β/α (i.e. with sum of squared deviations 2β, where deviations are from known mean μ)	$t_{2\alpha'}(\tilde{x}\|\mu, \sigma^2=\beta'/\alpha')$[4]
Normal with known mean μ	σ^2 (variance)	Scaled inverse chi-squared	$\nu,\ \sigma_0^2$	$\nu+n,\ \frac{\nu\sigma_0^2+\sum_{i=1}^n (x_i-\mu)^2}{\nu+n}$	variance was estimated from ν observations with sample variance σ_0^2	$t_{\nu'}(\tilde{x}\|\mu, \sigma_0^{2\prime})$[4]
Normal with known mean μ	τ (precision)	Gamma	$\alpha,\ \beta$ [note 4]	$\alpha+\frac{n}{2},\ \beta+\frac{\sum_{i=1}^n (x_i-\mu)^2}{2}$	precision was estimated from 2α observations with sample variance β/α (i.e. with sum of squared deviations 2β, where deviations are from known mean μ)	$t_{2\alpha'}(\tilde{x}\mid\mu, \sigma^2=\beta'/\alpha')$[4]
Normal[note 7]	μ and σ^2 Assuming exchangeability	Normal-inverse gamma	$\mu_0,\ \nu,\ \alpha,\ \beta$	$\frac{\nu\mu_0+n\bar{x}}{\nu+n},\ \nu+n,\ \alpha+\frac{n}{2},$ $\beta+\frac{1}{2}\sum_{i=1}^n(x_i-\bar{x})^2+\frac{n\nu}{\nu+n}\frac{(\bar{x}-\mu_0)^2}{2}$ • $\bar{x}$ is the sample mean	mean was estimated from ν observations with sample mean μ_0; variance was estimated from 2α observations with sample mean μ_0 and sum of squared deviations 2β	$t_{2\alpha'}\left(\tilde{x}\mid\mu', \frac{\beta'(\nu'+1)}{\nu'\alpha'}\right)$[4]
Normal	μ and τ Assuming exchangeability	Normal-gamma	$\mu_0,\ \nu,\ \alpha,\ \beta$	$\frac{\nu\mu_0+n\bar{x}}{\nu+n},\ \nu+n,\ \alpha+\frac{n}{2},$ $\beta+\frac{1}{2}\sum_{i=1}^n(x_i-\bar{x})^2+\frac{n\nu}{\nu+n}\frac{(\bar{x}-\mu_0)^2}{2}$ • $\bar{x}$ is the sample mean	mean was estimated from ν observations with sample mean μ_0, and precision was estimated from 2α observations with sample mean μ_0 and sum of squared deviations 2β	$t_{2\alpha'}\left(\tilde{x}\mid\mu', \frac{\beta'(\nu'+1)}{\alpha'\nu'}\right)$[4]

참고자료 : Wikipedia – conjugate prior

Basic Bayesian Terms

알려진 형태가 아닌 경우...

$$f(\theta|x) = \frac{p(\theta,x)}{p(x)} = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta} = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta} \propto p(x|\theta)p(\theta)$$

- 적분을 깔끔하게 수행할 수 없어, 수치적 (numerical) 방법들을 통해 구해야함

  1. 적분을 근사 (Laplace Approximation)

  2. 사후분포를 따르는 샘플을 추출 - 가능도와 사전분포를 이용해서

- 일반적으로 가장 많이 사용하는 방법은 MCMC (Markov Chain Monte Carlo) 기반 샘플링

  - 기초 평균/확률 모형 및 선형모형까진 정해진 꼴이 있지만, 로지스틱 회귀같은 간단한 모형도 어려움

Basic Bayesian Terms

Causal-Lab

MCMC 샘플링

- 사후분포의 범위를 포함하는 난수를 추출해서,

  뽑혀진 샘플이 사후분포를 얼마나 따르는지에 따라 확률적으로 채택/기각 (ex. Metropolis-Hastings)
- 하지만 기초적인 MCMC 알고리즘은 속도가 너무 느려서, 이를 위해 고려할게 많음
- Hamiltonian Monte Carlo (HMC) : 물리학 헤밀토니안 역학을 통해… 효율적인 샘플링을 지원함
- No-U-Turn Sampler (NUTS) : HMC를 개선. 대부분의 Bayesian 추론 툴의 기본으로 채택

Basic Bayesian Terms

Causal-Lab

MCMC 샘플링을 자세하게 알아야하나요?

- 과거 베이지안 추론이 잘 쓰이기 어려웠던 이유

  1. 샘플링 업데이트 구조를 직접 지정해줘야함 (수식 유도 필요)

  2. 샘플링을 위한 효율적인 알고리즘을 선택해야함

  3. 샘플링에 드는 속도 / 메모리 이슈

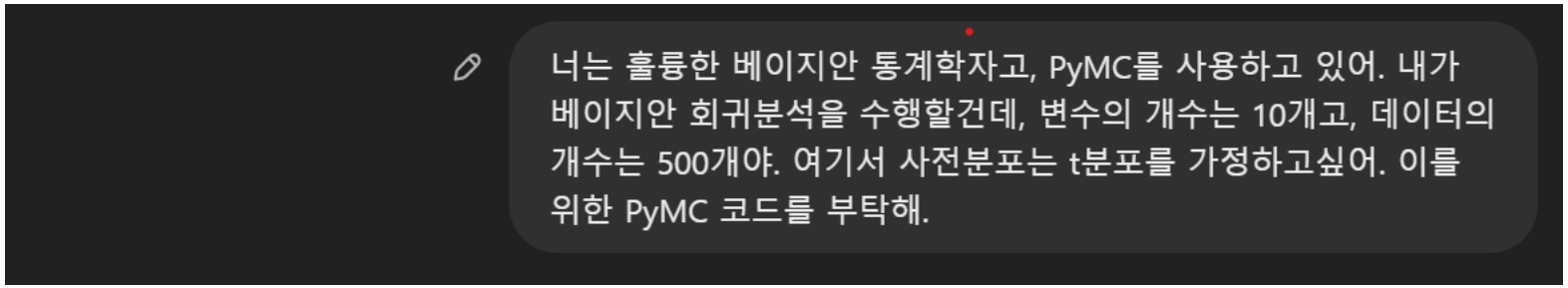
  4. 상용 오픈소스 솔루션의 허들 (JAGS, BUGS, STAN)

Basic Bayesian Terms

Causal-Lab

MCMC 샘플링을 자세하게 알아야하나요?

- 이젠 기본적인건 PyMC / Numpyro를 통해 다 지원
- 모형의 구조만 명확하게 지시할 수 있으면 문제 없음
- 직접적으로 업데이트 구조를 지정하지 않고, 최적의 세팅으로, 빠르게 계산
- 특히 ChatGPT 등과 호환이 매우 잘됨.



Basic Bayesian Terms

Causal-Lab

MCMC 샘플링을 잘 몰라도 됩니다

```python
# PyMC 모델 정의
with pm.Model() as model:
    # 사전분포: t-분포
    df = pm.Exponential("df", 1)  # 자유도는 양수여야 하므로 지수분포 사용
    beta = pm.StudentT("beta", nu=df, mu=0, sigma=1, shape=K)  # t-분포 사전
    sigma = pm.HalfStudentT("sigma", nu=3, sigma=1)  # 오차항의 표준편차도 t-분포
    mu = pm.math.dot(X, beta)

    # 종속변수의 likelihood
    y_obs = pm.Normal("y_obs", mu=mu, sigma=sigma, observed=y)

    # MCMC 샘플링
    trace = pm.sample(2000, tune=1000, target_accept=0.9)

# 결과 시각화
az.plot_trace(trace, var_names=["beta"])
az.summary(trace, var_names=["beta"])
```

Most Harmless Bayesian Prior

Causal-Lab

베이지안은 사전분포 때문에...

- 사전지식을 사전분포를 통해 반영한다? => 주관적이라는 걱정
- 일반적으로 베이지안에 거부감을 갖는 이유
- 실제 각 도메인 (의료 / 사회과학 등) 안에서는 테스크에 맞는 주관적 (subjective prior)를 사용
  - ex. 효과에 대한 사전분포로 N(4,1) 사용 / a,b,c 중에 선택에 대한 확률에 (0.8, 0.1, 0.1) 부여
- 객관적 (objective) 사전분포는 있는가?

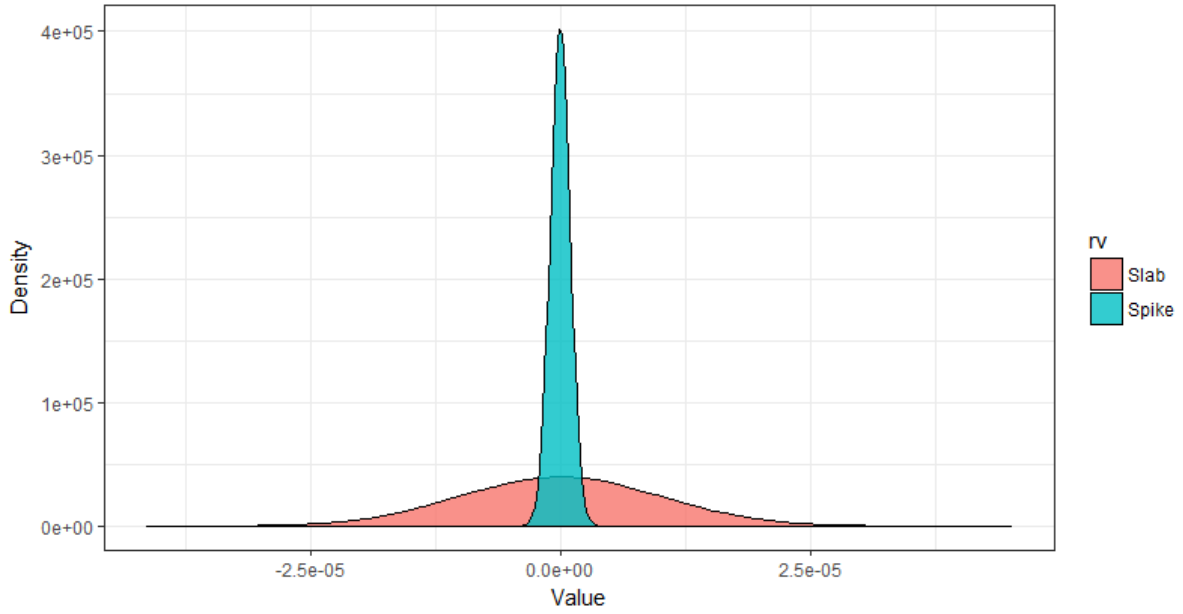
사전분포는 사전지식을 주입한다는 표현이 일단 맞을까?

Most Harmless Bayesian Prior

모형 구조에 대한 주입

- Ex. 베이지안 변수 선택

  - 유의미한 변수는 slab으로, 불필요한 변수는 spike를 통해 0으로 처리



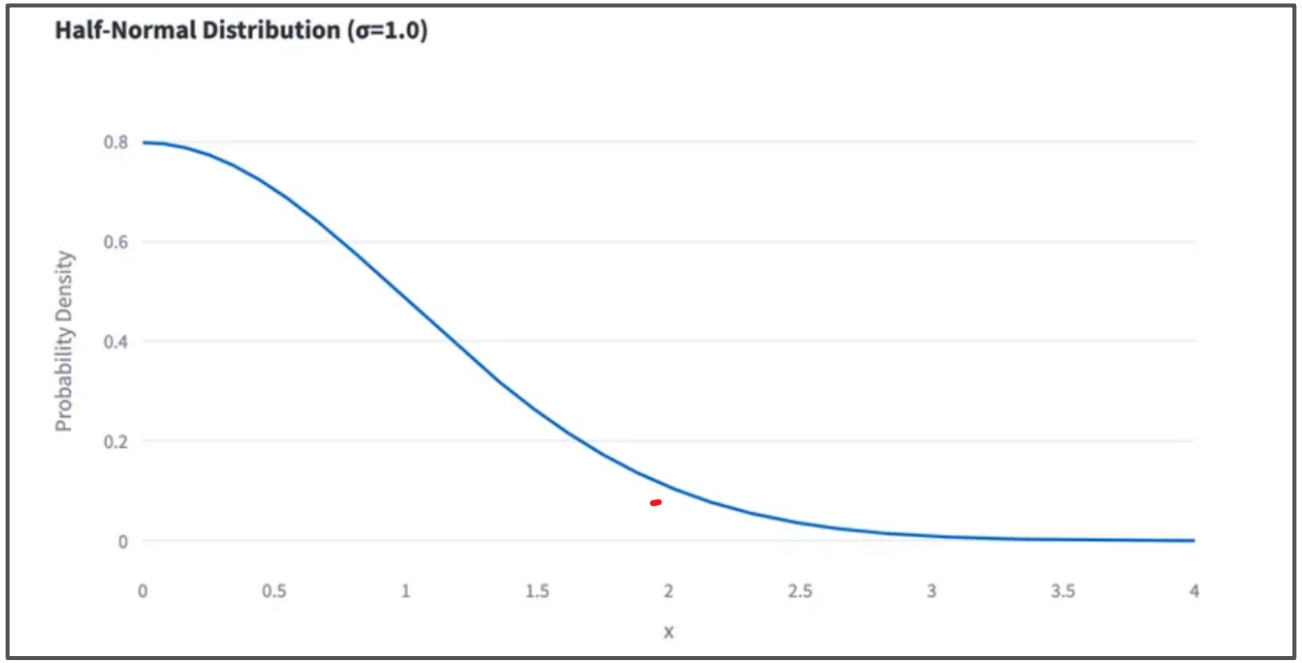
참고자료 : SPIKE AND SLAB VARIABLE SELECTION: FREQUENTIST AND BAYESIAN STRATEGIES (Ishwaran et al. 2005)

Most Harmless Bayesian Prior

모형 구조에 대한 주입

- Ex. 베이지안 MMM

  - 지출에 따른 음수 효과를 막기 위해 0 이상의 범위를 지니는 분포 사용



참고자료 : *Setting Priors in Bayesian Marketing Mix Modeling (MMM)*

Most Harmless Bayesian Prior

Causal-Lab

그래도 꺼림칙한데...

- 사전분포의 범위가 충분히 넓으면, 빈도주의 추론과 유사한 결론이 나온다는 사실이 알려져 있음

  ex. Bernstein-von Mises Thm - 데이터가 많은 경우 기본 정규모형과 완전히 일치된다

$$\left\| P(\theta|X) - N\left(\hat{\theta}, n^{-1}I(\theta_0)\right) \right\|_{TV} \xrightarrow{P_{\theta_0}} 0$$

- 기본적인 케이스들에서 사전분포가 데이터 몇 개의 영향력을 지니는지 계량가능

  - 일반적으로 권장되는 무정보적 (uninformative) 사전분포는 데이터 1개 이하의 영향력을 지님

데이터 (likelihood)가 사전분포 (prior)를 압도한다

참고자료 : Wikipedia - Bernstein-von Mises Theorem

Uncertantiy Quantification

Causal-Lab

Bayesian MMM

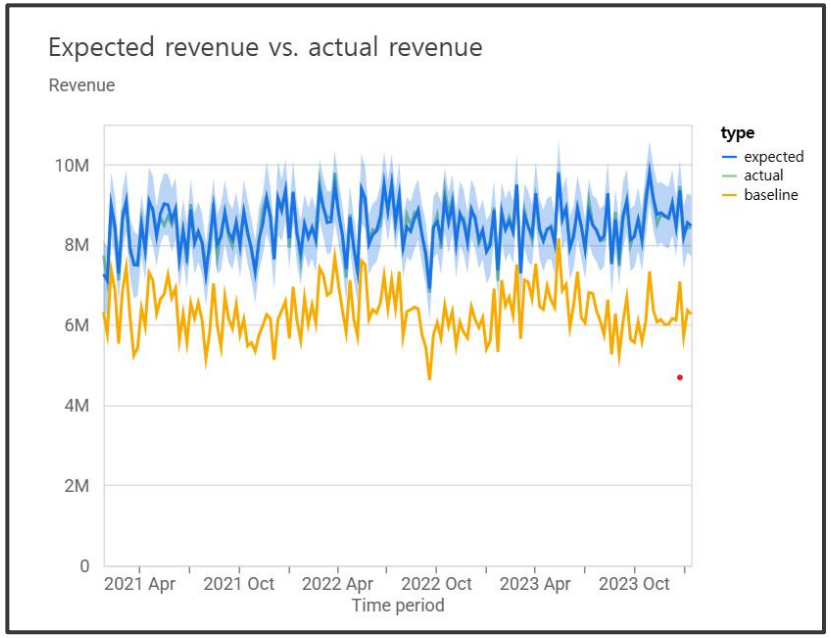
- MMM (Marketing Mix Modeling)은 예산 집행에 따른 효과를 추정하기 위한 모형
  - 성과 추적이 어려운 마케팅 상황에서 통계적 모델링을 통해 성과 배분
  - ex. TV 광고 / 옥외광고 / 개인정보보호 (iOS) 상황 등에 예산 비중이 높은 경우
- 이를 위한 모형으로 Meta Robyn, Google Lightweight MMM, Meridian 등 오픈소스 모형 존재
- 구글의 Google Lightweight MMM / Meridian MMM 은 베이지안 기반 방법론

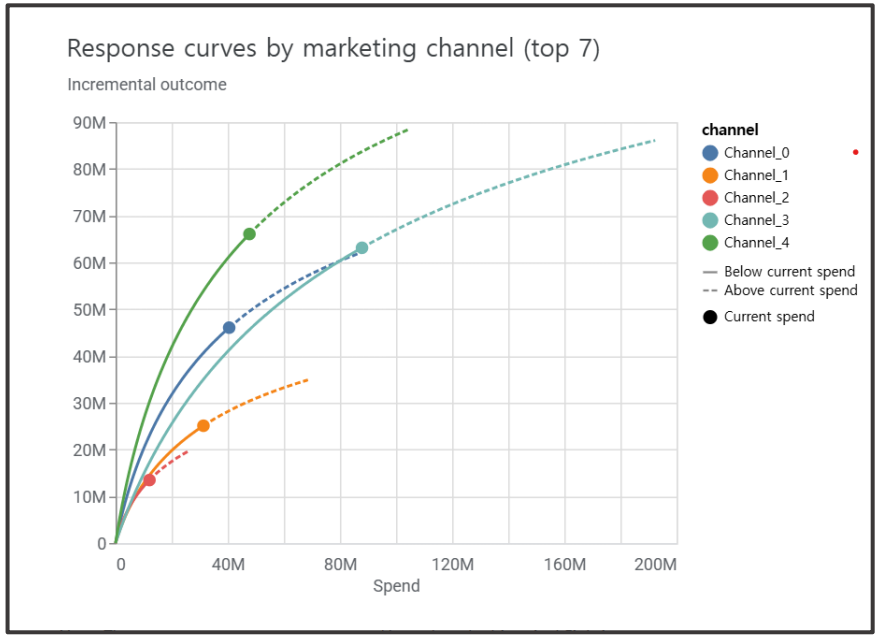
참고자료 : Google Lightweight MMM

Uncertantiy Quantification

Bayesian MMM

- 지출에 따른 매출 모델링 결과를 불확실성과 함께 확인 가능



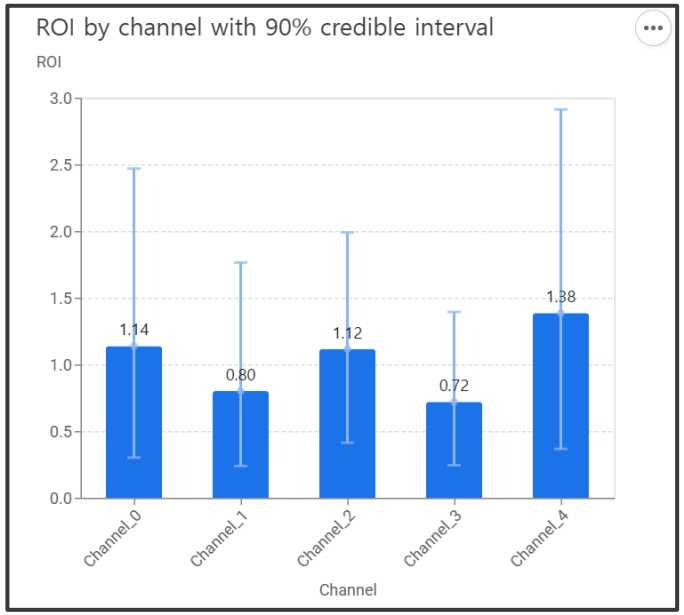


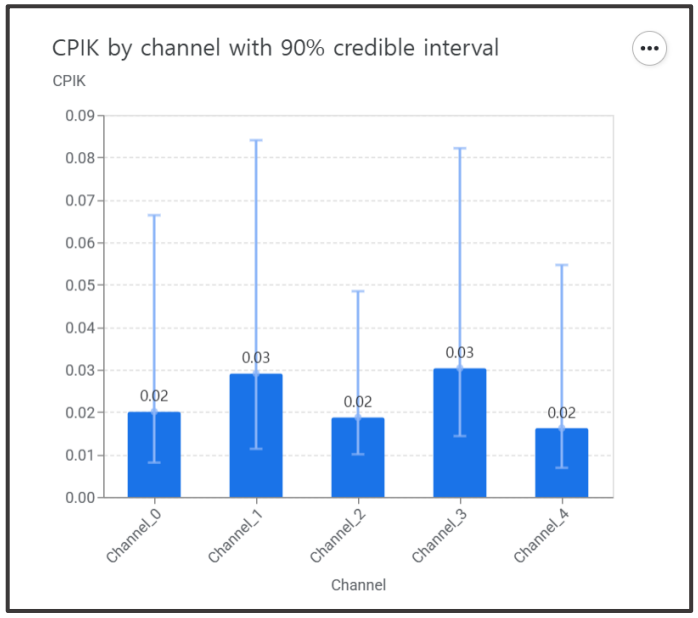
참고자료 : Google Meridian MMM

Uncertantiy Quantification

Bayesian MMM

- MMM의 관심은 단순 Y 예측이 아닌, 매체별 지출에 따른 증분
- 특정 KPI에 대한 불확실성이 너무 크다면 혼란을 야기할 수 있음





참고자료 : Google Meridian MMM

SKAN for iOS

- iOS의 경우 ATT (앱 추적 투명성)에 따라 개인정보추적 동의를 하지 않으면 광고성과 추적 불가

  - 이를 위한 대안 1) ATT 동의율 높이기, 2) SKAN 기반 광고성과로 피봇

- SKAN은 애플의 개인정보보호를 고려한 광고 솔루션

  - 기존 어트리뷰션과 별개로 작동하고, 유저레벨의 결과를 보여주지 않음

  - 일 / 매체 / 캠페인 유저 설치수 + 64비트 (케이스)로 집계된 레벨의 24시간 동안의 전환값을 전달

  - 전환값에는 구매 및 인앱 이벤트로 구성

    - (미구매 + 구매 7구간 = 8) * (인앱이벤트 3개 => 2^3=8) = 64

참고자료 : Appsflyer - SKAN

SKAN for iOS

- SKAN 매출구간 설정시, 해당 구간 유저수 * 구간의 중앙값을 통해 집계된 유입 후 24시간 추정 매출 반환

  - 최고 매출 한계를 넘어서거나, 군중익명성 조건을 만족하지 못하면 Null 전환값 반환

구간 번호	매출 시작값	매출 종료값	매출 중앙값	유저수	집계된 매출
1	1	3,000	1,500	2	3,000
2	3,000	10,000	6,500	2	13,000
3	10,000	20,000	15,000	4	60,000
4	20,000	35,000	27,500	2	55,000
5	35,000	60,000	47,500	2	95,000
6	60,000	100,000	80,000	2	160,000
7	100,000	200,000	150,000	1	150,000
8	200,000	inf	0	5	0
	Null	***Null***	***Null***	***5***	***0***

참고자료 : Appsflyer – SKAN Modeling

Hierarchical Modeling

Causal-Lab

SKAN for iOS

- SKAN 매출구간 모델링 결과에 따라, 24시간 Paid 매출의 추정치인 SKAN 매출의 신뢰도가 변함
- Appsflyer에서는 이를 “베이지안 계층모형”을 통해 모델링해, null 전환값의 매출을 추정

번호	매출 시작값	매출 종료값	매출 중앙값	유저수	집계된 매출
1	1	3,000	1,500	2	3,000
2	3,000	10,000	6,500	2	13,000
3	10,000	20,000	15,000	4	60,000
4	20,000	35,000	27,500	2	55,000
5	35,000	60,000	47,500	2	95,000
6	60,000	100,000	80,000	2	160,000
7	100,000	200,000	150,000	1	150,000
8	200,000	inf	0	5	0
	Null	***Null***	***Null***	***5***	***0***

Null 전환값이 어떤 구간에 속할지를 할당하는 문제

참고자료 : Appsflyer – SKAN Modeling

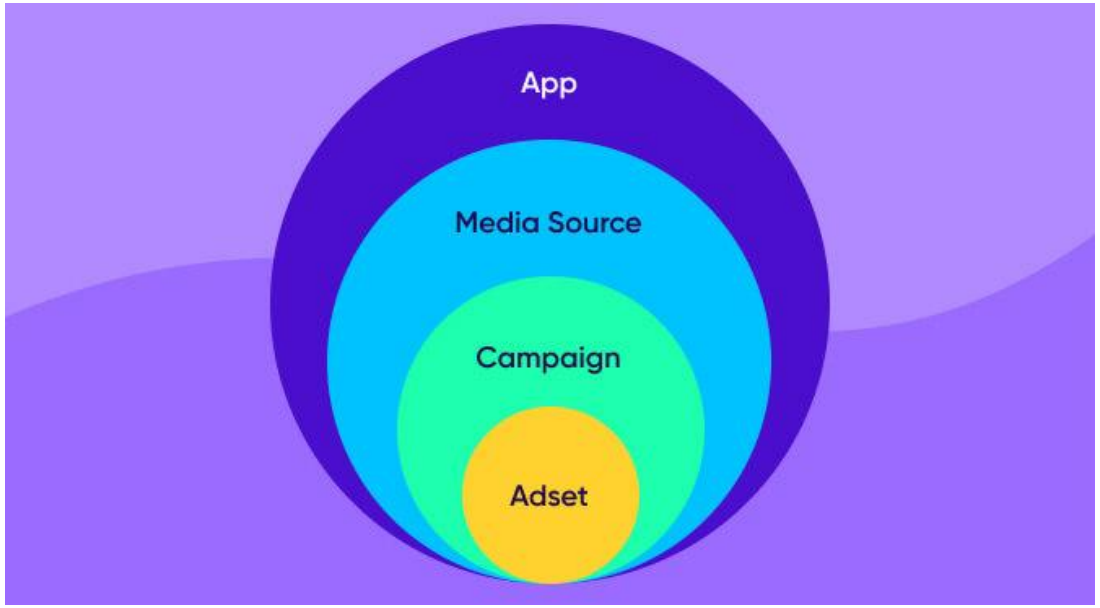
Hierarchical Modeling

SKAN for iOS

- Why Bayesian?

  개별 매체 / 캠페인 안에서만 분포를 고려할 경우 데이터가 너무 적어서,

  매체 / 캠페인간 정보들을 종합하기 위함



참고자료 : Appsflyer – SKAN Modeling

Hierarchical Modeling

Causal-Lab

SKAN for iOS

- Appsflyer 방식의 몇가지 이슈는 존재함

  - 전환값 Null 발생은 완전 무작위가 아니고, 고매출 구간에 발생할 확률이 더 높음

  - 전환값 Null 비율이 전체에서 20%를 넘어가면 추정에 문제 발생 가능

- SKAN 과 관련된 다음의 모델링이 가능하다고 알려져 있음

  - SKAN 적정 매출 구간 설정 모델링 (분포 추정)

  - SKAN과 독립적인 iOS Paid D1 매출에 대한 추정치 개발 (정답을 모르는 상황의 예측)

참고자료 : Appsflyer – SKAN Modeling

Online Updating + Interpretation

Causal-Lab

Frequentist A/B Test

- 그로스를 위해 수많은 A/B 테스트가 실행되고 있음

  - 상대적으로 실험을 수행하기 어려운/불가능한 영역들도 존재

- 상대적으로 인과에 가까운 추론이 가능하기 때문에, 프로덕트를 개선시킬 수 있다는 믿음 존재

- 전형적 A/B Test의 몇가지 한계

  - 결과 해석의 직관성 (p-value)

  - 실험 중간에 p-value가 임계치를 넘었다고 중단 (peaking)하는 것은 권장되지 않음

  - 샘플수를 검정력에 따라 미리 정의하는 것이 필요할 수 있음

참고자료 : PAP 이민호님 블로그 - 왜 군이 A/B 테스트인가요?

Online Updating + Interpretation

Bayesian A/B Test

- 많은 자료들은 Bayesian A/B Test 자체의 수행에 초점
- Bayesian A/B Test가 기존 A/B Test 대비 유용할 수 있는 점
  - 기대 손실(이익)을 고려한 모델링 통합이 자연스러움
  - 확률적 해석이 자연스러움
  - 실험 중간에 결과를 보고 종료 (peaking)하는 것이 가능 (반론 존재)
- 하지만 Bayesian A/B Test 를 도입하는 것엔 많은 장벽
  - 굳이? 리스크는? 실험 문화 측면? 진짜 더 나은 Test인가?

참고자료 : dynamicyield

그래서 결론은…?

Causal-Lab

All models are wrong, but some are useful.

가짜연구소

Causal-Lab

감사합니다.