한국금융학회·한국금융정보학회
공동 정책심포지엄 (2023.6.22)

데이터 프라이싱: 데이터 규모와 품질

임종호

연세대학교 통계데이터사이언스 학과

발표내용

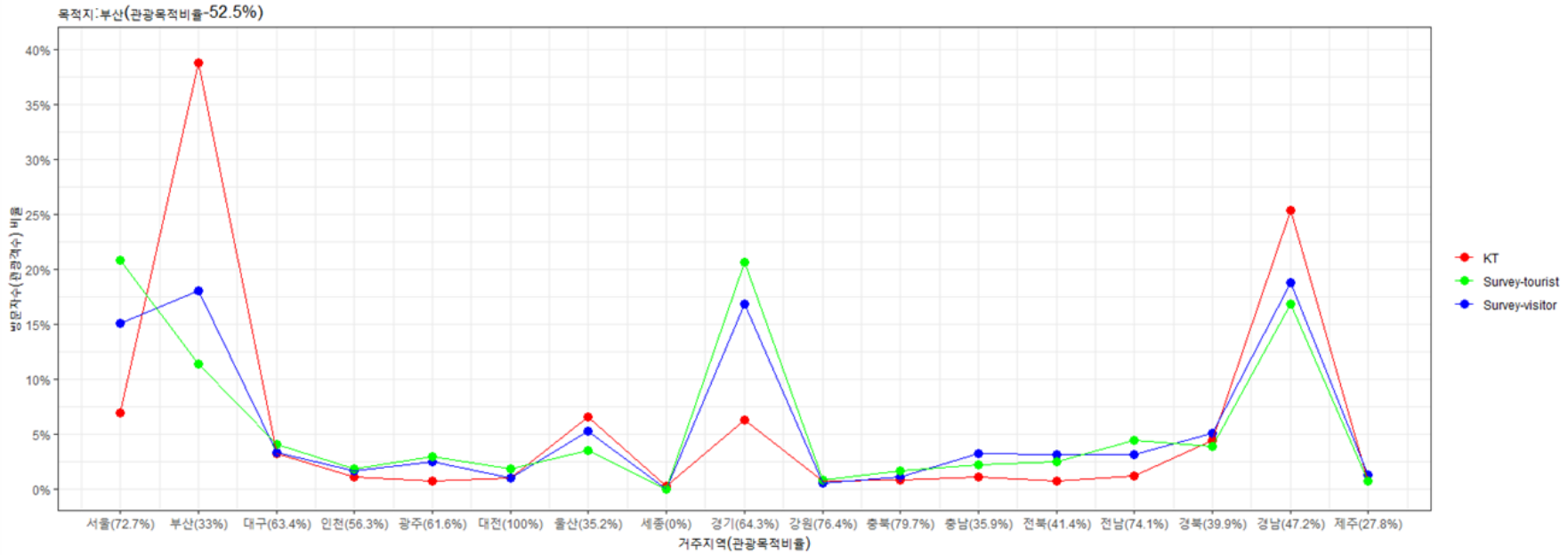
I 문제제기

II 데이터 외부성

III 데이터 편의

III 결론 및 시사점

I 연간 관광객 수 예제



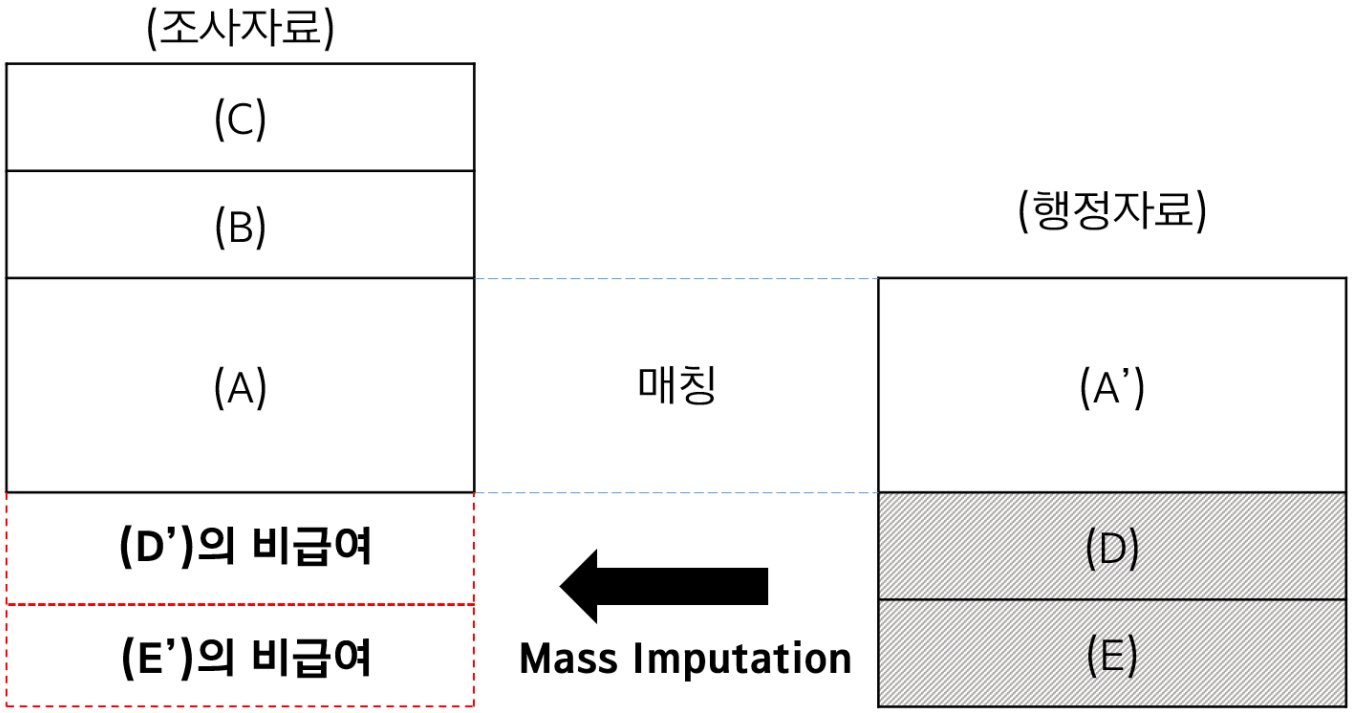
- 한국관광공사. (2019). “국민여행실태조사”와 KT 빅데이터를 활용한 추정 개선 방법론 연구
- 부산시에 방문한 방문자들의 지역별 통계
- 이동통신사 기준 방문자, 국민실태조사 방문자, 관광객 구분

I 연간 관광객 수 예제

거주지	방문자비율_KT	방문자비율_조사	보정방문자비율	95%신뢰구간_하한	95%신뢰구간_상한	KT_포함여부	방문자수_KT	보정방문자수_KT
서울	11.3%	18.4%	19.9%	13.4%	16.7%	0	9,869,390	17,305,437
대구	5.2%	4.1%	4.3%	3.1%	4.9%	1	4,529,873	3,748,925
인천	1.8%	2.0%	1.6%	1.0%	2.9%	1	1,596,927	1,405,523
광주	1.1%	3.0%	3.4%	1.7%	4.4%	1	987,122	2,994,237
대전	1.6%	1.2%	1.4%	0.5%	2.3%	1	1,419,602	1,248,936
울산	10.8%	6.4%	3.7%	5.2%	5.8%	0	9,370,583	3,247,326
세종	0.4%	0.0%	0.3%	0.0%	0.4%	1	324,099	285,614
경기	10.3%	20.6%	19.1%	17.2%	14.7%	0	9,002,265	16,608,907
강원	1.2%	0.7%	1.0%	0.2%	1.7%	1	1,032,072	908,203
충북	1.3%	1.3%	1.2%	0.6%	2.0%	1	1,165,237	1,025,956
충남	1.8%	4.0%	4.0%	2.5%	5.8%	1	1,527,810	3,453,186
전북	1.2%	3.9%	3.5%	2.2%	4.8%	1	1,071,501	3,010,568
전남	2.0%	3.8%	3.6%	2.5%	4.5%	1	1,720,880	3,137,526
경북	7.2%	6.2%	6.4%	5.0%	8.4%	1	6,291,473	5,536,907
경남	41.5%	22.9%	25.5%	20.4%	34.2%	1	36,120,118	22,241,841
제주	1.2%	1.6%	1.1%	0.8%	2.1%	1	1,087,199	957,058

- 서울시에서 “부산”을 방문한 비율이 11.3%에서 19.9%로 증가
- 조사데이터를 벤치마킹으로 사용했기에 보정비율은 조사데이터와 유사한 것을 확인
- 조사데이터의 품질에 따라서 보정 결과값의 품질도 결정됨

Ⅰ 행정자료 활용 예제



- (행정자료) 급여정보는 있으나 비급여 정보는 없음
- 한국의료패널(조사자료)에서 비급여 정보 조사
- 자료연계를 통하여 급여+비급여 정보 결합

고민해 볼 거리

- 빅데이터, 행정자료는 우리가 필요로 하는 정보를 주는가?
- 데이터 외부성(externality)가 데이터 가격(작성 비용)에 어떠한 영향을 주는가?
- 개인/민감 정보를 포함하는 경우, 데이터 규모와 품질은 어떻게 되는가?
- 개인/민감 정보에 대한 가격시스템이 정보 제공 여부와 관련이 있는 경우 데이터 품질 및 가격이 어떻게 되는가?
- 데이터 관련 정책 및 행정자료 활용에 주는 시사점?

목표: Toy 데이터 모형을 통하여 최대한 답해보기

II 모형 셋업

$$\begin{pmatrix} Y \\ X \\ x \end{pmatrix} \sim N\left(\begin{bmatrix} \beta_0 + \beta_1 \mu_x \\ \mu_x \\ \mu_x \end{bmatrix}, \begin{bmatrix} \beta_1^2 \sigma_{xx} + \sigma_{ee} & \beta_1 \sigma_{xx} + \sigma_{eu} & \beta_1 \sigma_{xx} \\ \beta_1 \sigma_{xx} + \sigma_{eu} & \sigma_{xx} + \sigma_{uu} & \sigma_{xx} \\ \beta_1 \sigma_{xx} & \sigma_{xx} & \sigma_{xx} \end{bmatrix} \right)$$

Acemoglu et al. "Too Much Data: Prices and Inefficiencies in Data Markets" 참고

1. $x \sim N(\mu_x, \sigma_{xx})$: unobserved target value
2. $X = x + u$ with $u \sim N(0, \sigma_{uu})$: observable value
3. $Y = \beta_0 + \beta_1 x + e$ with $e \sim N(0, \sigma_{ee})$: external information:

   - 대상의 추가정보(예: 행정자료, 마이데이터)나 유사 대상의 정보 등

4. (가정) $\beta_0 = 0$ & $\texttt{cov}(e, u) \equiv \sigma_{eu} = 0$ (0일 필요는 없음)
5. 예측값-$\hat{x}$ 예측오차-$\hat{x} - x$
6. 목표: 예측오차의 분산을 최소화 $\min_{\hat{x}} \texttt{var}(\hat{x} - x)$

II (시나리오 1) No Information

활용가능한 정보가 없는 경우 최적의 예측값:

$$\hat{x} \equiv E(x|\cdot) = \mu_x.$$

- 예측오차: $\mu_x - x$.
- 예측오차의 분산

$$v_0 \equiv \texttt{var}(\hat{x} - x) = \sigma_{xx}.$$

II (시나리오 2) X 관측

X가 관측(수집)된 경우, 최적의 예측값:

$$\begin{aligned}\hat{x} \equiv E(x|X) &= \mu_x + \frac{\sigma_{xx}}{\sigma_{xx}+\sigma_{uu}}(X-\mu_x) \\ &= \frac{\sigma_{xx}}{\sigma_{xx}+\sigma_{uu}}X + \frac{\sigma_{uu}}{\sigma_{xx}+\sigma_{uu}}\mu_x\end{aligned}$$

- 예측오차

$$(X-x) - \frac{\sigma_{uu}}{\sigma_{xx}+\sigma_{uu}}X + \frac{\sigma_{uu}}{\sigma_{xx}+\sigma_{uu}}\mu_x.$$

- 예측오차의 분산

$$v_X \equiv \mathtt{var}(\hat{x}-x) = \sigma_{uu} - \frac{\sigma_{uu}^2}{\sigma_{xx}+\sigma_{uu}}$$

- $v_0 \geq v_X$: X가 예측오차의 분산을 줄여줌 $\left(\sigma_{xx} \quad vs \quad \sigma_{uu} - \frac{\sigma_{uu}^2}{\sigma_{xx}+\sigma_{uu}}\right)$

II (시나리오 3) Y 관측

Y가 관측(수집)된 경우, 최적의 예측값:

$$\begin{aligned}\hat{x} \equiv E(x|Y) &= \mu_x + \frac{\beta_1 \sigma_{xx}}{\beta_1^2 \sigma_{xx} + \sigma_{ee}}(Y - \mu_y) \\ &= \frac{\beta_1 \sigma_{xx}}{\beta_1^2 \sigma_{xx} + \sigma_{ee}} Y + \frac{\sigma_{ee}}{\beta_1^2 \sigma_{xx} + \sigma_{ee}} \mu_x\end{aligned}$$

- 예측오차

$$\frac{\beta_1 \sigma_{xx}}{\beta_1^2 \sigma_{xx} + \sigma_{ee}} Y + \frac{\sigma_{ee}}{\beta_1^2 \sigma_{xx} + \sigma_{ee}} \mu_x - x$$

- 예측오차의 분산

$$v_Y \equiv \mathtt{var}(\hat{x} - x) = \sigma_{xx} - \frac{\beta_1^2 \sigma_{xx}^2}{\beta_1^2 \sigma_{xx} + \sigma_{ee}}$$

- $v_0 \geq v_Y$: Y가 예측오차의 분산을 줄여줌 $\left(\sigma_{xx} \quad vs \quad \sigma_{xx} - \frac{\beta_1^2 \sigma_{xx}^2}{\beta_1^2 \sigma_{xx} + \sigma_{ee}}\right)$

II (시나리오 4) X와Y 모두 관측

X와 Y가 모두 관측 가능할 때 최적의 예측값

$$\hat{x} \equiv E(x|Y,X) \quad = \quad \gamma_0 + \gamma_1 Y + \gamma_2 X$$

where $\gamma_0 = (1-\gamma_2)\mu_x - \gamma_1 \mu_y$ and

$$(\gamma_1, \gamma_2)^\top \quad = \quad \Sigma_{ZZ}^{-1} \begin{pmatrix} \beta_1 \sigma_{xx} \\ \sigma_{xx} \end{pmatrix}$$

where $Z = (Y, X)$ and Σ_{ZZ} is the variance of Z.

- 예측오차

$$\gamma_0 + \gamma_1 Y + \gamma_2 X - x$$

- 예측오차의 분산

$$v_{X,Y} \equiv \mathtt{var}(\hat{x} - x) \quad = \quad \sigma_{uu} - \frac{\beta_1^2(\sigma_{xx} + \sigma_{ee})\sigma_{uu}^2}{\beta_1^2 \sigma_{xx}\sigma_{uu} + \sigma_{ee}(\sigma_{xx} + \sigma_{uu})}$$

- $v_0 \geq v_{X,Y}$: X와 Y 정보가 예측오차의 분산을 줄여줌

II 소결

1. $v_Y \geq v_X \quad \text{if} \quad \sigma_{uu} \leq \frac{\sigma_{ee}}{\beta_1^2}$

2. $v_X \geq v_{X,Y}$

- 외부정보가 있으면 예측에 대한 gain이 발생함 (Externality)

  p: data price/cost, v: price of privacy
  v와 p의 크기에 따라서 X를 직접적으로 관측하는 것에 대한 요인이 작아짐 (Acemoglu et al.)

- 정보가 많을수록 gain이 커짐

- 데이터 품질에 따라서 gain의 크기가 달라짐

III 선택편의가 존재하는 경우

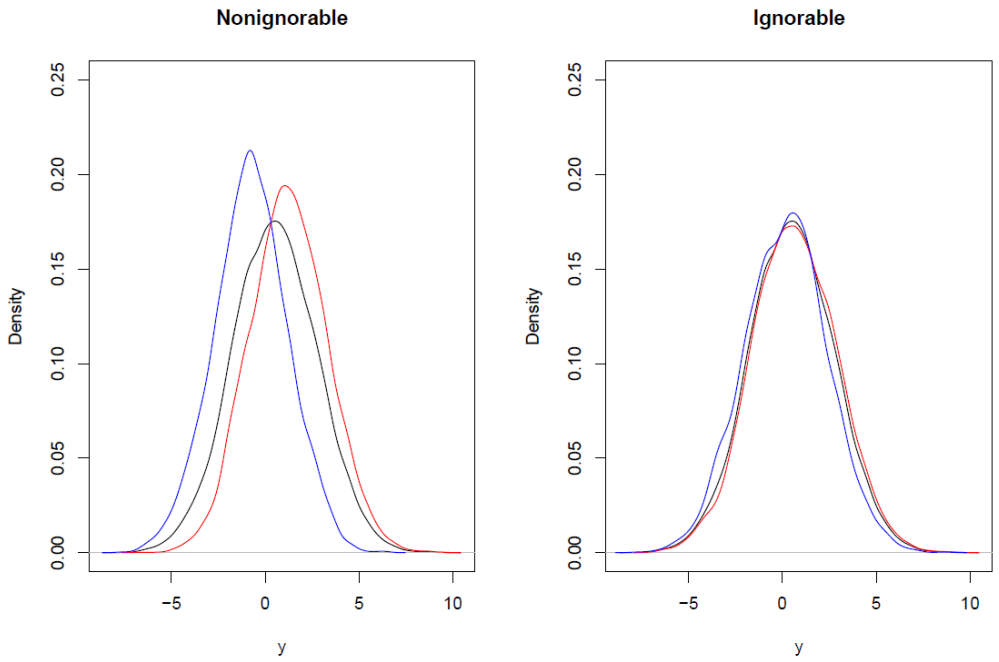
- R: X 관측 여부에 대한 지시함수
- 데이터 가격/비용이 동일하다고 하면, X관측여부를 간단히 모델링 할 수 있음

$$\begin{aligned} \mathtt{pr}(R = 1 \mid \nu) &= \{1 + \exp(-\phi_0 - \phi_1 \nu)\}^{-1} \\ &= \{1 + \exp(-\phi_0 - \phi_1 x)\}^{-1} \end{aligned}$$

- 정보수집 여부가 프라이버시 가치에 의존한다고 가정
- 논의의 편의상 프라이버시 가치는 true outcome가 동일하다고 가정 ($\nu = x$)

  예) 소득이 높은 사람은 본인 소득정보에 대한 가치를 더 높게 생각

III Nonignorable 메커니즘



- Nonignorable mechanism

$$f(x \mid X, R = 1) \propto f(x \mid X)\mathtt{pr}(R = 1 \mid x, X)$$

$\mathtt{pr}(R = 1 \mid x, X)$가 x에 의존하기 때문에

$$E(x \mid X) \neq E(x \mid X, R = 1) \neq E(x \mid x, R = 0)$$

III 선택편의 비용

- $x \mid (X, R = 1) \sim N(\alpha_0 + \alpha_1 X, \sigma_r^2)$ 가정
- 응답 모형

$$\mathtt{pr}(R = 1 \mid \nu) \quad = \quad \{1 + \exp(-\phi_0 - \phi_1 x)\}^{-1}$$

- Tukey's representation

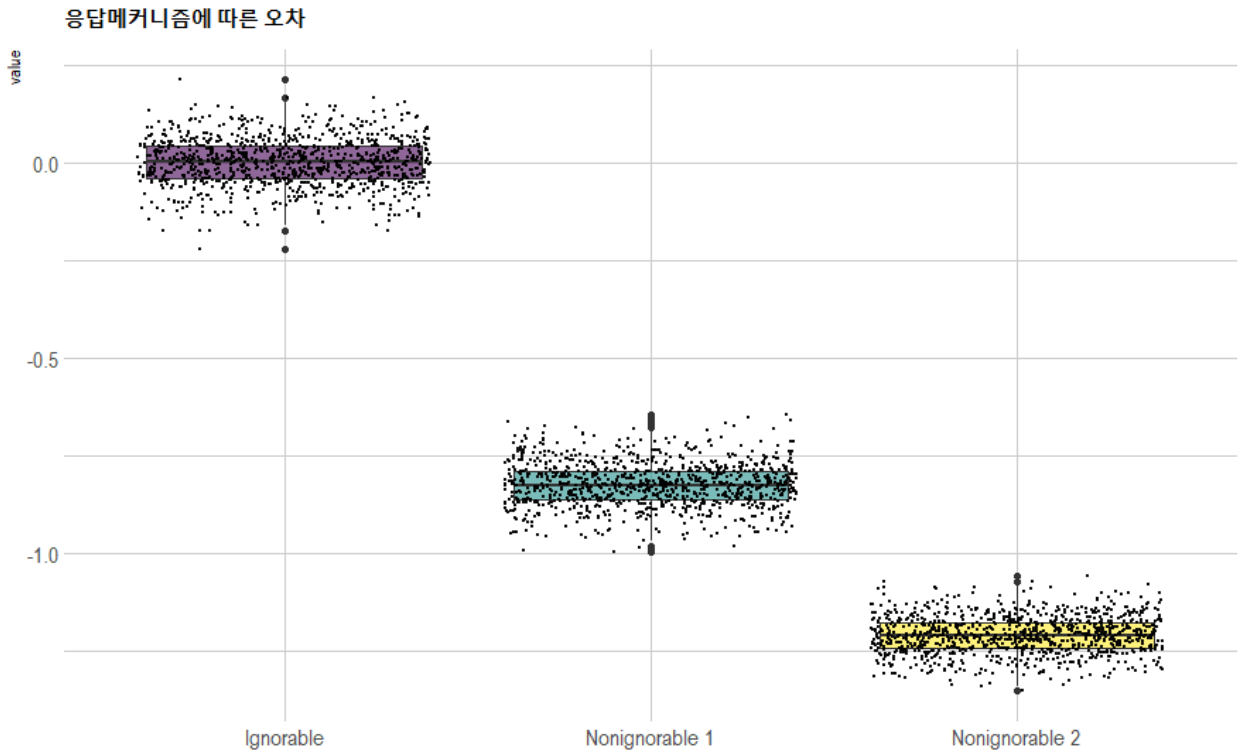
$$f(x \mid X, R = 0) \propto x \mid (X, R = 1) \frac{\mathtt{pr}(R = 0 \mid x)}{\mathtt{pr}(R = 1 \mid x)}$$

- $x \mid (X, R = 0) \sim N(\alpha_0 + \alpha_1 X - \phi_1 \sigma_r^2)$
- 예측값 차이가 발생 / 선택편의에 따른 비용(오차)이 발생

$$E(x \mid X, R = 1) - E(x \mid X, R = 0) = \phi_1 \sigma_r^2$$

- 규모에 대한 가격(비용)보다 품질에 따른 가격(비용)이 더 클 수 있음

III 모의실험



- E(x | X, R=1) - E(x | X, R=0)을 추정
- Ignorable: 랜덤하게 X 관측
- Nonignorable 1: 선택편의 크기 φ=-1
- Nonignorable 2: 선택편의 크기 φ=-2

III 소결

- 데이터 품질은 데이터 자체에 대한 품질(예: 측정오차)뿐만 아니라 수집/작성되는 메커니즘(예: 선택편의)과도 밀접하게 연계되어 있음

- 대부분의 관측데이터는 응답 메커니즘에 노출되어 있으며, 이 때 발생한 데이터 편향은 무시할 수 없는 데이터 가격(비용)이 될 수 있음

- 응답 메커니즘에 따라서 발생하는 추가비용은 데이터 수집의 물리적 비용에 비례하지 않음 → 데이터 품질에 따른 가격 차별화가 어려움

IV 결론 및 시사점

- 데이터 externality가 데이터의 가격(비용)을 줄이는 역할을 함

- 개인정보 가치 크기가 정보 제공 여부(응답 메커니즘)에 영향을 주는 경우 데이터 규모와 가격(비용)을 모두 높이는 기제가 됨

- 응답 메커니즘으로 인한 정보 왜곡은 데이터 규모 증대로 해결될 수 없음에도 대부분의 경우 데이터 규모의 문제로 접근하고 있음
  (예: 행정자료-이동통신사, 주택가격동향지수)

- 무분별한 행정자료의 사용은 의도하지 않은 더 큰 비용을 지불해야 하는 상황을 만들 수 있음

감사합니다!!