Linear Model in Multidimensional Space

Interpret *U*-shaped relationship as Linear

Jinseob Kim[1], Joohon Sung[1,*]

Dec 20, 2017

Introduction

Motivation: Curved Light

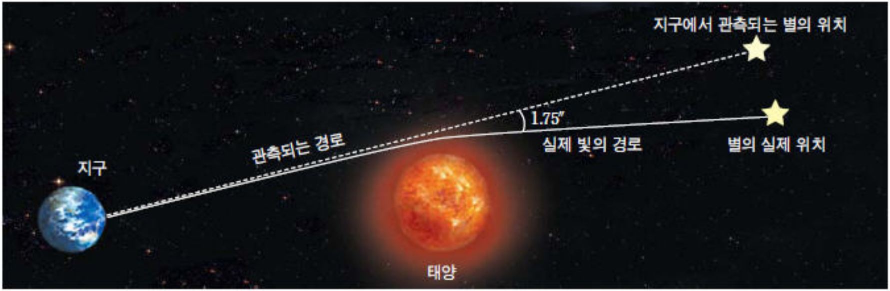
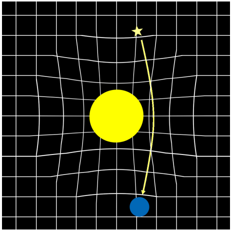


그림 I－60 태양 주위에서 빛의 휨

- 빛이 휜다?
- 뉴턴: 태양의 중력이 빛을 끌어당긴다.
- But, 빛의 질량은 0.

Einstein's General relativity: Curved Spacetime



- 빛은 직선이 맞다, 주변 **(시)공간이 휘어진 것**이다.
- 3차원 공간 $\to$ 휘어진 4차원 시공간

$$g_{uv} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} g_{tt} & g_{tx} & g_{ty} & g_{tz} \\ g_{tx} & g_{xx} & g_{xy} & g_{zx} \\ g_{ty} & g_{xy} & g_{yy} & g_{yz} \\ g_{tz} & g_{zx} & g_{yz} & g_{zz} \end{pmatrix}$$

Non-linear Issues: *U*-shape

U-shape relationship은 가장 흔한 non-linear issue[1–4].

1. Linear model 그대로 사용.
2. Non-linear model[5,6]

- Threshold : 2 parameter
- Square, cube: 2∼3 parameter
- GAM: 1∼10 parameter?
- Neural Network: 10? 100?

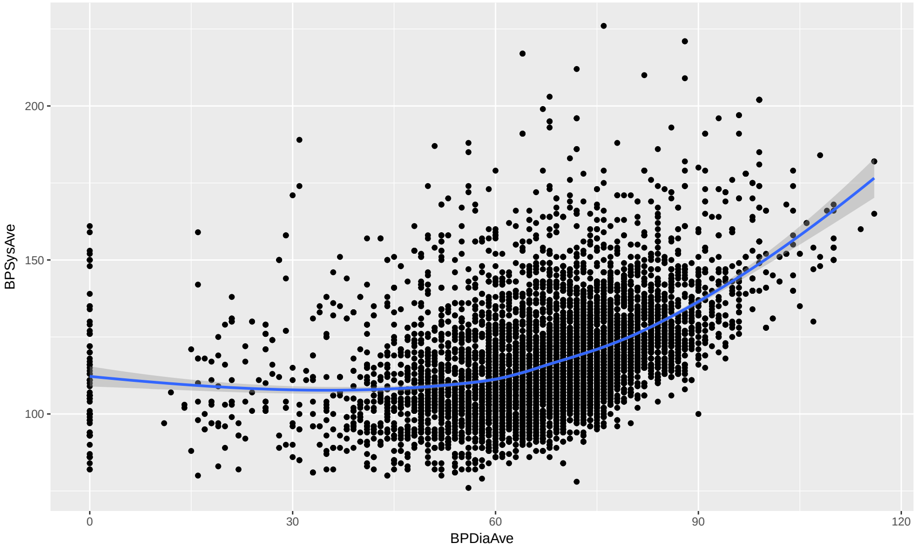


Figure 1: GAM

Simple is best.

- Linear Model의 장점: 설명하기 쉽다.
  - 변수당 1 parameter
- Non-linear model은 휘어진 모양을 해석
  - 설명이 복잡

Main Topic

Multi-Dimensional Linear Model(MDLM)

- **휘어진 다차원공간으로 선형모형 확장**.

관계가 비선형(X), 주변공간이 휘어짐(O)

- 선 $\rightarrow$ 면, 곡면, 공간. . .
- 새로운 무대에서는 선형관계.

Contents

- 기존 선형모형을 완전히 포함한 개념 설계.
- Simulation을 통해 기존 모형들과 비교.
- 실제 ER data에 적용

Formula

Generalization: 2 variables, 2 dimensions

$$\vec{\boldsymbol{Y}} = (\beta_{01} + \beta_1 X_1)\vec{\boldsymbol{g}}_1 + (\beta_{02} + \beta_2 X_2)\vec{\boldsymbol{g}}_2$$

- $\vec{\boldsymbol{g}}_{\boldsymbol{i}}$: 단위벡터

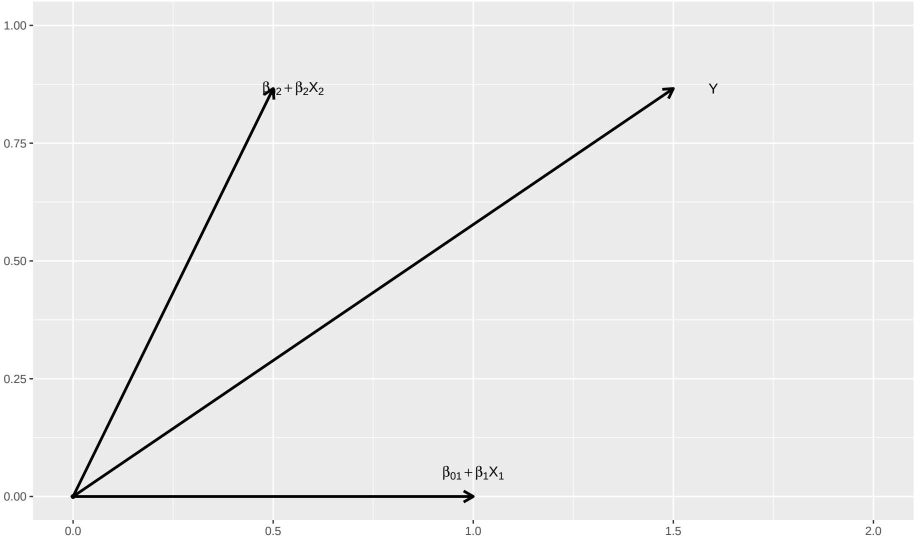


Figure 2: $cos\theta = g_{12}$

Interpretation: Linear!!

$$\vec{\boldsymbol{Y}} = (\beta_{01} + \beta_1 X_1)\vec{\boldsymbol{g}}_1 + (\beta_{02} + \beta_2 X_2)\vec{\boldsymbol{g}}_2$$

$$\begin{aligned} d\vec{\boldsymbol{Y}} &= \beta_1 dX_1 \vec{\boldsymbol{g}}_1 + \beta_2 dX_2 \vec{\boldsymbol{g}}_2 \\ &= \beta_1 d\vec{\boldsymbol{X}}_{\boldsymbol{1}} + \beta_2 d\vec{\boldsymbol{X}}_{\boldsymbol{2}} \end{aligned}$$

* X_2가 고정되었을 때, $\vec{\boldsymbol{Y}}$는 $\vec{\boldsymbol{X}}_{\boldsymbol{1}}$의 방향으로 β_1만큼 증가한다.

Generalization of Linear Model

- If $\vec{g}_1 = \vec{g}_2$
    - $g_{12} = 1$

$$\begin{aligned} Y &= (\beta_{01} + \beta_1 X_1) + (\beta_{02} + \beta_2 X_2) \\ &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 \end{aligned}$$

- Same to linear model

Scala version

$$Y^2 = (\beta_{01} + \beta_1 X_1)^2 + (\beta_{02} + \beta_2 X_2)^2 + 2g_{12}(\beta_{01} + \beta_1 X_1)(\beta_{02} + \beta_2 X_2)$$

- $\vec{\boldsymbol{g}}_{\mathbf{1}} \cdot \vec{\boldsymbol{g}}_{\mathbf{2}} = g_{12}\ (0 \le g_{12} \le 1)$

$$Y^2 = (\beta_{01} + \beta_1 X_1 + g_{12}(\beta_{02} + \beta_2 X_2))^2 + (1 - g_{12}^2)(\beta_{02} + \beta_2 X_2)^2$$

- $X_1 = -\frac{\beta_{01} + g_{12}(\beta_{02} + \beta_2 X_2)}{\beta_1}$에서 최소값을 갖는 U-shape

Generalization: p variables, 2 dimensions

$$\vec{\boldsymbol{Y}} = (\beta_{01} + \beta_1 X_1 + \cdots + \beta_l X_l)\vec{\boldsymbol{g}}_1 + (\beta_{02} + \beta_{l+1} X_{l+1} \cdots + \beta_p X_p)\vec{\boldsymbol{g}}_2$$

$$Y^2 = (\beta_{01} + \beta_1 X_1 + \cdots + \beta_l X_l)^2 + (\beta_{02} + \beta_{l+1} X_{l+1} \cdots + \beta_p X_p)^2$$
$$+ 2g_{12}(\beta_{01} + \beta_1 X_1 + \cdots + \beta_l X_l)(\beta_{02} + \beta_{l+1} X_{l+1} \cdots + \beta_p X_p)$$

Generalization: p variables, p dimensions

$$\vec{\boldsymbol{Y}} = (\beta_{01} + \beta_1 X_1)\vec{\boldsymbol{g}}_1 + (\beta_{02} + \beta_2 X_2)\vec{\boldsymbol{g}}_2 + \cdots (\beta_{0p} + \beta_p X_p)\vec{\boldsymbol{g}}_p$$
$$= \sum_{i=1}^{p} (\beta_{0i} + \beta_i X_i)\vec{\boldsymbol{g}}_i$$

$$Y^2 = \sum_{i=1}^{p} (\beta_{0i} + \beta_i X_i)\vec{\boldsymbol{g}}_i \cdot \sum_{i=1}^{p} (\beta_{0i} + \beta_i X_i)\vec{\boldsymbol{g}}_i$$
$$= \sum_{i=1}^{p} (\beta_{0i} + \beta_i X_i)^2 + 2\sum_{i<j} g_{ij}(\beta_{0i} + \beta_i X_i)(\beta_{0j} + \beta_j X_j)$$

Estimation

Least Square method

$$SSE(\beta) = \sum_{k=1}^{N} \left(Y_k - \sqrt{\sum_{i=1}^{n} (\beta_i X_{ki} + \beta_{i0})^2 + 2 \sum_{i<j} g_{ij} (\beta_i X_{ki} + \beta_{i0})(\beta_j X_{kj} + \beta_{j0})}\right)^2$$

- If all $g_{ij} = 1$
  - 기존 선형모형의 최소제곱추정법과 동일
  - 자연스러운 일반화

Optimization

- No analytical solution.
- Various optimization methods[7–9].
- **optim** & **constrOptim** function in *R*

P value calculation

- hessian matrix(H) : SSE를 두번 미분한 값[10].

$$SSE(\hat{\theta} + d\theta) = SSE(\hat{\theta}) + H \cdot \frac{(d\theta)^2}{2}$$

$$(d\theta)^2 = 2 \cdot H^{-1} \cdot (SSE(\hat{\theta} + d\theta) - SSE(\hat{\theta}))$$

- Generalization[10,11]

$$\text{vcov}(\hat{\beta}) = 2 \cdot H^{-1} \cdot MSE(\hat{\beta})$$

Curved Space: Fixed vs from Data

1. Fixed space 지정

$$g_{ij} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

2. Data에서 직접 추정

$$\begin{pmatrix} 1 & g_{12} & g_{13} \\ g_{21} & 1 & g_{23} \\ g_{31} & g_{32} & 1 \end{pmatrix}$$

Estimation of g_{ij}

- β들과 g_{ij}들을 같이 추정.
- GEE(Generalized Estimating Equation)와 비슷
  - Working correlation matrix를 직접 구할 수 있음[12].
- g_{ij}: 0에서 1사이의 제한조건
  - constrained optimization technique[13].

Simulation

Compare Model

- X_1, X_2: (1,1), (1,2),$\cdots$, (1,10), (2,1), $\cdots$, (10,10)

1. Linear Model
2. MDLM (1): fixed $g_{12} = 0$
3. MDLM (2): estimation g_{12} from data
4. Polynomial(Quadratic) Model
5. GAM[14]

Scenario 1: $Y = X_1 + X_2$

- Sampling $Y \sim N(X_1 + X_2, 1)$

	Linear	MDLM (1)	MDLM (2)	Quadratic	GAM
RMSE	1 ± 0	1.3 ± 0	1 ± 0	1 ± 0	1 ± 0.1
DF	4	5	6	6	5.3 ± 1.4
AIC	286.3 ± 8.8	343.2 ± 4.1	288.5 ± 9.6	288.4 ± 9	284.1 ± 11.1

Scenario 2: $Y^2 = X_1^2 + X_2^2$

- Sampling $Y \sim N(\sqrt{X_1^2 + X_2^2}, 1)$

	Linear	MDLM (1)	MDLM (2)	Quadratic	GAM
RMSE	1.1 ± 0	1 ± 0	1 ± 0	1.1 ± 0	1.1 ± 0.1
DF	4	5	6	6	6.5 ± 0.9
AIC	314.4 ± 4	285.4 ± 4.4	287.4 ± 4.4	310.8 ± 9.2	308 ± 9

Scenario 3: $\vec{Y} = (\beta_{01} + \beta_1 X_1)\vec{g}_1 + (\beta_{02} + \beta_2 X_2)\vec{g}_2$

- Sampling $Y \sim N(\sqrt{(\beta_{01} + \beta_1 X_1)^2 + (\beta_{02} + \beta_2 X_2)^2 + 2g_{12}(\beta_{01} + \beta_1 X_1)(\beta_{02} + \beta_2 X_2)}, 1)$

	Linear	MDLM (1)	MDLM (2)	Quadratic	GAM
RMSE	1.2 ± 0.1	1.1 ± 0.1	1 ± 0	1.1 ± 0.1	1.1 ± 0.1
DF	4	5	6	6	5.9 ± 0.4
AIC	319.7 ± 17.7	311.3 ± 12	298.1 ± 3.5	314.4 ± 15.5	314.9 ± 15.6

Apply to Real Data

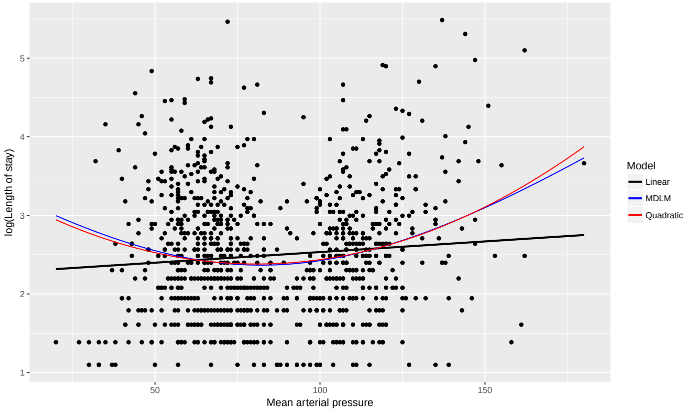
ER data

http://biostat.mc.vanderbilt.edu/dupontwd/wddtext/data/3.25.2.SUPPORT.csv

- 응급실 내원 당시 평균 동맥압(mean arterial pressure, MAP)과 재실기간(length of stay,LOS)
- log(LOS)를 **intercept와 MAP의 2차원**에서 표현.

$$\log(\vec{\text{LOS}}) = \beta_{00}\vec{\boldsymbol{g_1}} + (\beta_{01} + \beta_1 \cdot \text{MAP})\vec{\boldsymbol{g_2}}$$

map	intcpt	los	loglos
20	1	4	1.386294
27	1	4	1.386294
30	1	3	1.098612
30	1	4	1.386294



- Linear: $\log(\text{LOS}) = 2.2624 + 0.0027 \cdot \text{MAP}$ (AIC 2434)
- MDLM: $\log(\text{LOS})^2 = 2.3669^2 + (-2.4276 + 0.0295 \cdot \text{MAP})^2$ (AIC 2413)
- Quadratic: $\log(\text{LOS}) = 3.3742 - 0.0246 \cdot \text{MAP} + 2 \times 10^{-4} \cdot \text{MAP}^2$ (AIC 2414)

Discussion

의의

- 휘어진 다차원 공간에서 간단하게 U-shape을 해석
- Linear 컨셉 유지
  - X 하나당 parameter 1개
- 기존 선형모형을 완벽히 포함한 일반화
  - p값 계산 가능

활용

- Non fixed g_{ij}: U-shape 관계를 더 정밀하게 추정. 휘어진 공간에 대한 해석
- Fixed g_{ij}: 공간구조 고정(ex: 독립된 2차원)하여 직관적인 해석

GEE와 비교(1)

- GEE- independent

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

- Fixed g_{ij}=0

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

GEE와 비교(2)

- GEE: Compound Symmetry/Exchangeable

$$\begin{pmatrix} 1 & r & r & r \\ r & 1 & r & r \\ r & r & 1 & r \\ r & r & r & 1 \end{pmatrix}$$

- Fixed $g_{ij} = g$

$$\begin{pmatrix} 1 & 1 & g & g \\ 1 & 1 & g & g \\ g & g & 1 & 1 \\ g & g & 1 & 1 \end{pmatrix}$$

GEE와 비교(3)

- GEE: unstructured

$$\begin{pmatrix} 1 & r_{12} & r_{13} & r_{14} \\ r_{21} & 1 & r_{23} & r_{24} \\ r_{31} & r_{32} & 1 & r_{34} \\ r_{41} & r_{42} & r_{43} & 1 \end{pmatrix}$$

- Non fixed g_{ij}

$$\begin{pmatrix} 1 & g_{12} & g_{13} & g_{14} \\ g_{21} & 1 & g_{23} & g_{24} \\ g_{31} & g_{32} & 1 & g_{34} \\ g_{41} & g_{42} & g_{43} & 1 \end{pmatrix}$$

한계 (1): $Y \geq 0$ 만 다룰 수 있음.

$$
\begin{aligned}
Y^2 &= \sum_{i=1}^{p} (\beta_{0i} + \beta_i X_i)\vec{\boldsymbol{g}}_i \cdot \sum_{i=1}^{p} (\beta_{0i} + \beta_i X_i)\vec{\boldsymbol{g}}_i \\
&= \sum_{i=1}^{p} (\beta_{0i} + \beta_i X_i)^2 + 2 \sum_{i<j} g_{ij} (\beta_{0i} + \beta_i X_i)(\beta_{0j} + \beta_j X_j)
\end{aligned}
$$

- Health 연구에서 $Y < 0$ 인 경우는 거의 없음.
- $Y' = Y - Y_{min}$ 등 변수치환 활용.

Suggestion: Dirac's Idea[15]

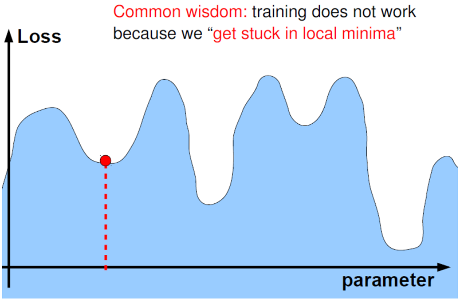
- *Paul Dirac*: 특수상대성이론을 고려한 양자역학의 방정식 Dirac Equation.
  - 방정식의 계수(β)가 꼭 숫자일 필요없다. 행렬이어도 됨.

$$\beta_0 = \alpha_0 \times \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}, \beta_1 = \alpha_1 \times \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix}$$

$$Y = \sqrt{\beta_0^2 + \beta_1^2 x_1^2 + \beta_2^2 x_2^2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

한계(2): Local minima issues

- β를 추정하는데 optimization technique를 사용함.
  - SSE(β)의 진짜 최소값(Global minimum) 이 아닐 수 있음.



- 최근 연구에서 고차원 공간인 경우 local minima problem은 매우 희귀한 것으로 나타났음.
  - 모든 차원에서 local minima일 가능성은 매우 낮기 때문[16]

Conclusion

- Einstein: 공간의 무대를 3차원이 아니라 휘어진 4차원으로 확장한다면 빛은 여전히 직선[17]

$$\nabla^2 \Phi = 4\pi G \rho_0 \rightarrow \boldsymbol{R}_{uv} - \frac{1}{2}\boldsymbol{g}_{uv} = \frac{8\pi G}{c^4}\boldsymbol{T}_{uv}$$

- 본 연구: **선형공간의 무대를 휘어진 다차원 공간으로 확장**하여 *U*-shape을 선형관계로 바라볼 수 있다.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \rightarrow \vec{\boldsymbol{Y}} = (\beta_{01} + \beta_1 X_1)\vec{\boldsymbol{g}}_1 + (\beta_{02} + \beta_2 X_2)\vec{\boldsymbol{g}}_2$$

기존 선형모형이 놓치는 건강관련 현상을 휘어진 다차원 변수공간에서 간단하게 설명할 수 있을 것이다.

References

1. Calabrese EJ, Baldwin LA. U-shaped dose-responses in biology, toxicology, and public health. *Annual review of public health* 2001; 22: 15–33.

2. Power C, Rodgers B, Hope S. U-shaped relation for alcohol consumption and health in early adulthood and implications for mortality. *The Lancet* 1998; 352: 877.

3. De Wit LM, Van Straten A, Van Herten M, et al. Depression and body mass index, a u-shaped association. *BMC public health* 2009; 9: 14.

4. Knutson KL, Turek FW. The u-shaped association between sleep and health: The 2 peaks do not mean the same thing. *Sleep* 2006; 29: 878–879.

5. Jagodzinski W, Weede E. Testing curvilinear propositions by polynomial regression with particular reference to the interpretation of standardized solutions. *Quality and Quantity* 1981; 15: 447–463.

6. Buja A, Hastie T, Tibshirani R. Linear smoothers and additive models. *The Annals of Statistics* 1989; 453–510.

7. Nelder JA, Mead R. A simplex method for function minimization. *The computer journal* 1965; 7: 308–313.

8. Fletcher R, Reeves CM. Function minimization by conjugate gradients. *The computer journal* 1964; 7: 149–154.

9. Byrd RH, Lu P, Nocedal J, et al. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing* 1995; 16: 1190–1208.

10. Richter P. Estimating errors in least-squares fitting. *The Telecommunications and Data Acquisition Report* 1995; 107–137.

11. Venables WN, Smith DM, Team RDC, et al. An introduction to r. 2004; 59–62.

12. Pan W, Connett JE. Selecting the working correlation structure in generalized estimating equations with application to the lung health study. *Statistica Sinica* 2002; 475–490.

13. Rios LM, Sahinidis NV. Derivative-free optimization: A review of algorithms and comparison of software implementations. *Journal of Global Optimization* 2013; 56: 1247–1293.

14. Wood SN. Mgcv: GAMs and generalized ridge regression for r.