清華大學
Tsinghua University
政治学系

Learning Latent Variable Analysis with Dr. Hu

潜在变量分析

胡悦
清华大学政治学系

内容概要

因素分析(Factorial Models)

1. 探索性影子分析(EFA)
2. 验证性因果分析(CFA)
3. 结构方程模型(SEM)

类型分析(Typological Models)

1. 项目反应理论(IRT)
2. 跨群组项目反应(MrP, GIRT, DCPO)

操作语言

- R[1]
  - `mirt`
  - `DCPO`

[1] 现存处理IRT的R packages已超过50个。

因子分析不香么?

1. 假定潜在变量是连续的;
2. 对于指标不区分变量类型;
3. 难以捕捉群组差异
4. EFA无法囊括指标间关系;
5. CFA面临“简略理论vs测量质量”的矛盾

IRT优势

1. 天生为二元指标设计（衍生适应定序变量和连续变量）；
2. 易与Bayesian inference结合，解决潜在变量scale不确定问题;
3. 在Bayesian框架下更好解决缺失值和“Don't Know”问题;
4. 易与跨群组估计结合，实现指标跨组可比

个人层级IRT

应用范围：社会调查

调查问题：

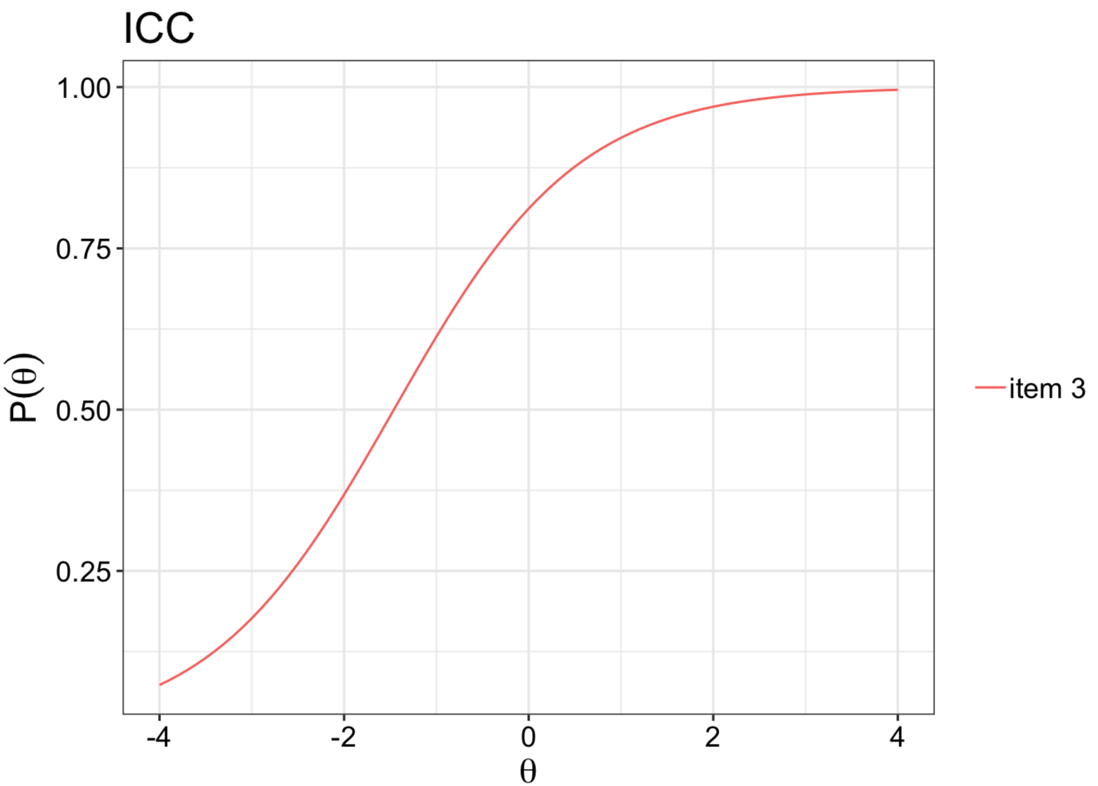
1. Yes/No
2. 可以转化为二元的问题
3. 定序问题（e.g., Liker scale questions）

IRT 假定

1. Monotonicity
2. Unidimensionality
3. Local independence
4. Parameter invariance

Monotonicity

单增趋势：随潜在变量增加，获得1的可能性也随之增加。



Unidimensionality

- 聚合的项目均指向同一个潜在变量。
- 基于理论

直到引入multidimensional IRT

Local Independency

对于每一项目（e.g.,一道题）的响应(e.g., 选择的选项)间的关联性只来自共同的潜在变量。

换言之，控制潜在变量影响后，问题间响应相互独立

$$P(y_{iq}, y_{i'q} \mid \theta_q) = P(y_{iq} \mid \theta_q) P(y_{i'q} \mid \theta_q)$$

Parameter Invarance

- Parameters在项目间不变
- Parameters在响应人群间不变[1]
  - 当进行Multiple Group IRT时尤可能被违反

[1] 通过基于Wald and likelihood-ratio approach来检测Differential item functioning (DIF).

Rasch Model (1PL)
→ Two-Parameter Logistic Model (2PL)
→ Three-Parameter Logistic Model (3PL)
→ Four-Parameter Logistic Model (4PL)

Group IRT

- $y_{iq} \in \{0,1\}$: subject `i`'s score on question `q`
- $\theta_i \in \{-\infty, +\infty\}$: Unbounded latent trait
- σ_q: Difficulty

$$\Pr(y_{iq} = 1) = \mathrm{logist}^{-1}(\theta_i - \sigma_q)$$

Item Response

Response Theory

操作案例 (Bock & Lieberman 1970)

Law School Admissions Test, sec 7

5个yes/no问题

Item.1	Item.2	Item.3	Item.4	Item.5
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

Difficulty Parameter

```r
m_lsat <- mirt(df_lsat, model = 1, itemtype = "Rasch", verbose
coef(m_lsat, simplify = TRUE) %>%
  kable(format = "html")
```

	a1	d	g	u
Item.1	1	1.8680718	0	1
Item.2	1	0.7909134	0	1
Item.3	1	1.4608233	0	1
Item.4	1	0.5214399	0	1
Item.5	1	1.9927710	0	1

	x
F1	0

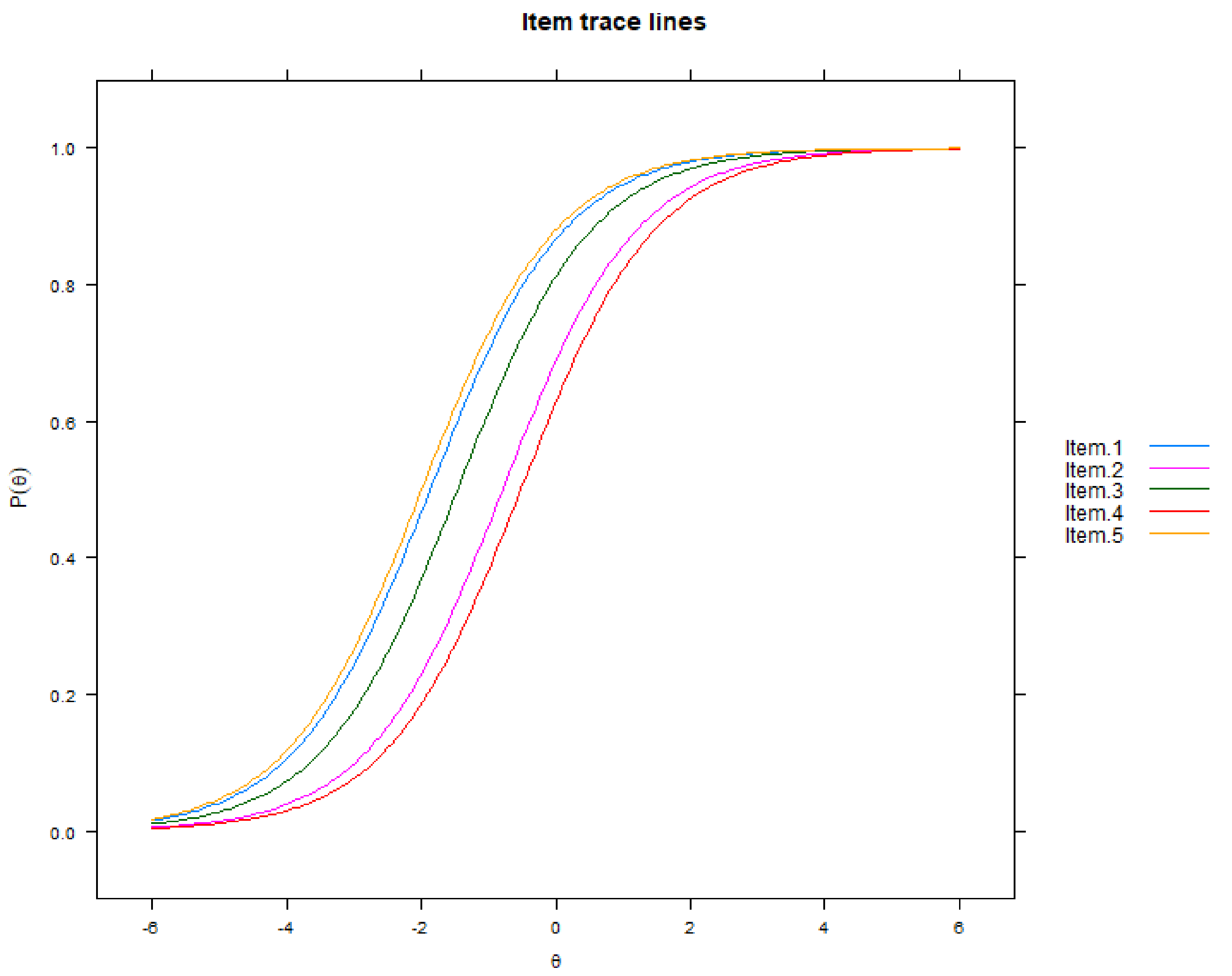
	F1
F1	1.021944

Please always diagnose your results,

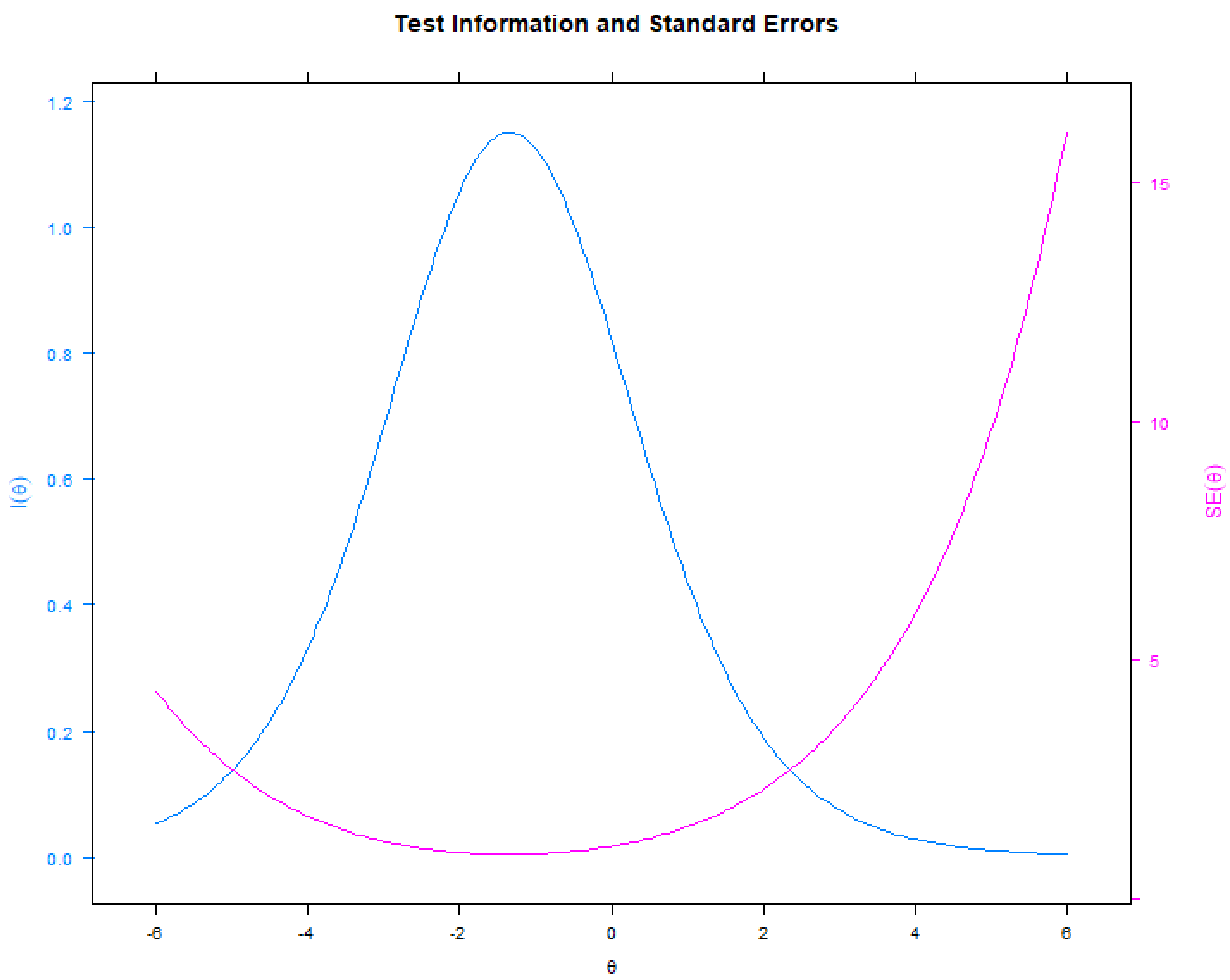
and **understand** what you are diagnosing.

--- Dr. Yue Hu

Item Characteristic Curves (ICC)



Test Characteristic Curve



Rasch局限：Measurement error

Two-Parameter Logistic Model (2PL IRT)

$$Pr(y_{iq} = 1) = logist^{-1}(\kappa_q \theta_i - \sigma_q)$$

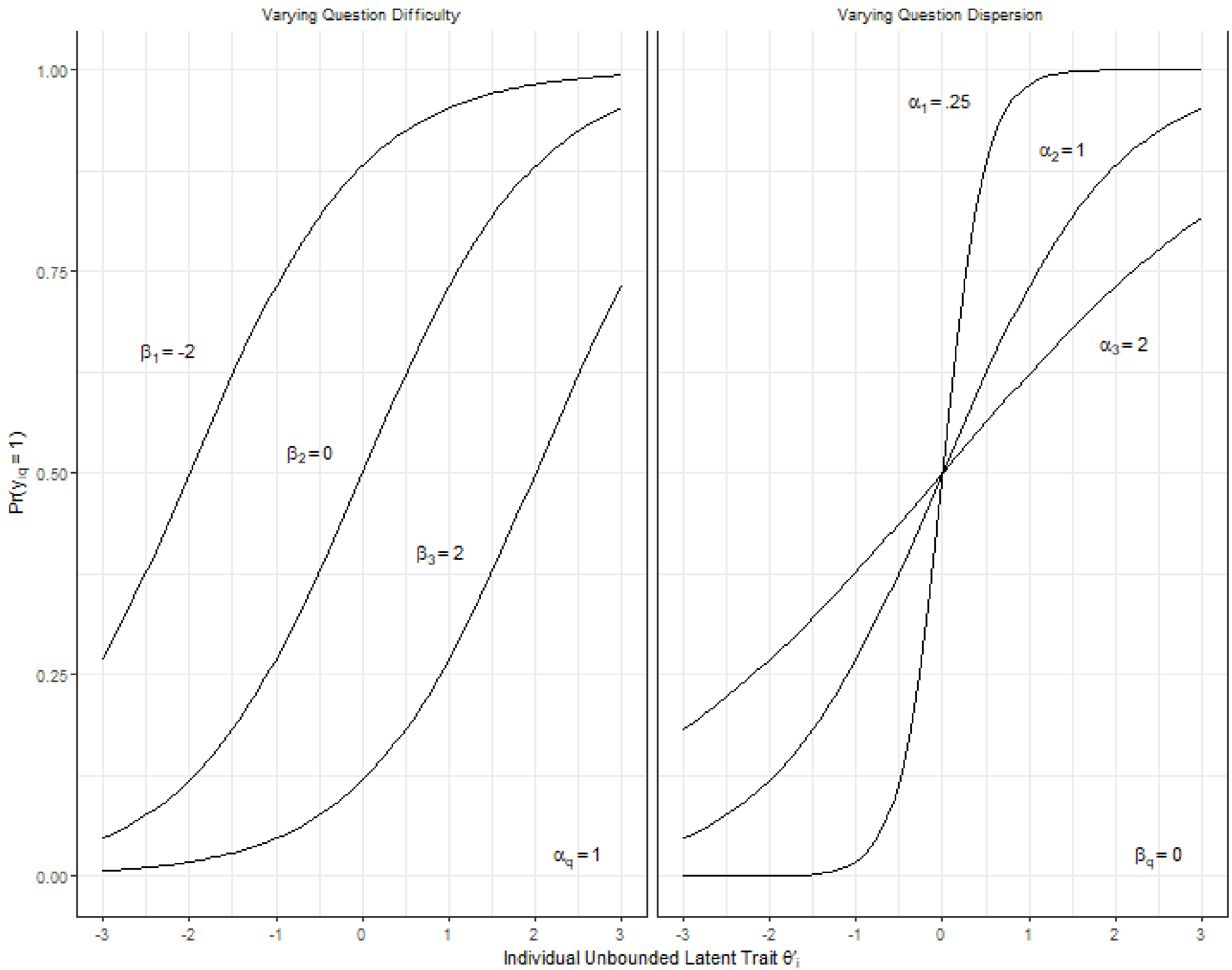
κ_q: Discrimination (Parameter of dispersion)

另一种常见写法

$$Pr(y_{iq} = 1) = logist^{-1}[\frac{\theta_i - \beta_q}{\alpha_q}]$$

β_q: σ_q / κ_q, threshold("difficulty", 控制location)

α_q: κ_q^{-1}, dispersion (控制斜率)


Varying Question Difficulty
Varying Question Dispersion
1.00
0.75
0.50
0.25
0.00
Pr(y_iq = 1)
β1 = -2
β2 = 0
β3 = 2
αq = 1
α1 = .25
α2 = 1
α3 = 2
βq = 0
-3
-2
-1
0
1
2
3
Individual Unbounded Latent Trait θ'i

```r
m_lsat2PL <-  mirt(df_lsat, model = 1, itemtype = "2PL", verbos
coef(m_lsat2PL, simplify = TRUE)
```

```
## $items
##            a1     d g u
## Item.1 0.988 1.856 0 1
## Item.2 1.081 0.808 0 1
## Item.3 1.706 1.804 0 1
## Item.4 0.765 0.486 0 1
## Item.5 0.736 1.855 0 1
##
## $means
## F1
##  0
##
## $cov
```

需要2PL吗？

Likelihood-Ratio Test

```
##
## Model 1: mirt(data = df_lsat, model = 1, itemtype = "Rasch",
## Model 2: mirt(data = df_lsat, model = 1, itemtype = "2PL", ve
```

AIC	SABIC	HQ	logLik	df	p
5341.802	5352.192	5352.994	-2664.901	NaN	NaN
5337.610	5354.927	5356.263	-2658.805	4	0.0159822

如果有人全凭猜咋办？——大量低θ人群

Three-Parameter Logistic Model (3PL)

$$Pr(y_{iq}=1) = c_i + (1-c_i)logist^{-1}[\frac{(\theta_i - \beta_q)}{\alpha_q}]$$

c_i：Item lower asymptote ("guessing")

极大增加演算成本→通常需要1000以上观测点

如果有人不care咋办

Four-Parameter Logistic Model (4PL)

$$Pr(y_{iq} = 1) = c_i + (d_i - c_i)logist^{-1}[\frac{(\theta_i - \beta_q)}{\alpha_q}]$$

d_i：Item upper asymptote ("carelessness"), d < 1

鉴于3PL已经需要1000-ish观测点……

IRT 统计检验

- 测试层：Global fit
- 项目层：Item fit & residual
- 个体层：Personal fit

Global Fit[1]

$$G^2 = 2[\sum_{l}^{s} r_l ln(\frac{r_l}{N\tilde{P}_l})]$$

N: 参与人数

l: 可能的反应

r: 做出特定反应的人数

当数据过于稀疏时(item > 10)，M2, M2*

```
M2(m_lsat)
```

```
##                M2 df           p      RMSEA    RMSEA_5   RMSEA_9!
## stats 23.17287  9 0.00581954 0.03970314 0.02003961 0.0599830:
##               SRMSR       TLI       CFI
## stats 0.04744033 0.9284234 0.9355811
```

[1] RMSEA, SRMSR, CFI, TLI对于IRT同样使用

Covariation-based residuals

```
residuals(m_lsat)
```

```
## LD matrix (lower triangle) and standardized values:
##
##         Item.1 Item.2 Item.3 Item.4 Item.5
## Item.1      NA -0.017  0.020  0.022  0.019
## Item.2  0.292     NA  0.105 -0.042 -0.064
## Item.3  0.389 10.976     NA  0.007  0.007
## Item.4  0.474  1.801  0.055     NA -0.052
## Item.5  0.362  4.063  0.045  2.691     NA
```

Single item/person fit

```
# Item
itemfit(m_lsat, fit_stats = "infit")
```

```
##      item outfit z.outfit infit z.infit
## 1 Item.1  0.744   -3.597 0.939  -1.025
## 2 Item.2  0.758   -7.500 0.826  -6.303
## 3 Item.3  0.711   -5.420 0.860  -3.202
## 4 Item.4  0.770   -8.877 0.818  -7.962
## 5 Item.5  0.797   -2.572 0.993  -0.081
```

```
# Person
personfit(m_lsat)
```

```
##         outfit   z.outfit     infit     z.infit         Zh
## 1    0.6420958 -0.9001784 0.6953214 -0.96202244  0.93429336
## 2    0.6420958 -0.9001784 0.6953214 -0.96202244  0.93429336
## 3    0.6420958 -0.9001784 0.6953214 -0.96202244  0.93429336
## 4    0.6420958 -0.9001784 0.6953214 -0.96202244  0.93429336
## 5    0.6420958 -0.9001784 0.6953214 -0.96202244  0.93429336
## 6    0.6420958 -0.9001784 0.6953214 -0.96202244  0.93429336
## 7    0.6420958 -0.9001784 0.6953214 -0.96202244  0.93429336
```

如果出现问题：

1. 通过S-χ^2、local dependency等检查观测和估计数值差别
2. 改变model type, 比如2PL → 3PL
3. 如果最初用binary，尝试polytomous或者nominal response models
4. 尝试non-parametric smoothing techniques

延展1：一维到多维

传统IRT：一维聚合

Multidimentional IRT (MIRT, Phil Chalmers, 2015)

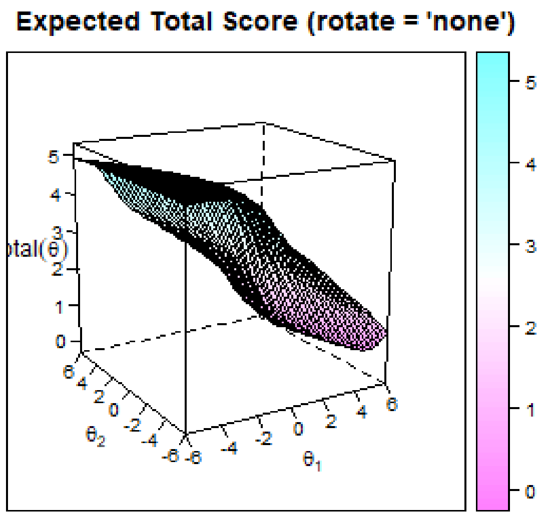
$$Pr(y_{iq}=1) = logist^{-1}[\frac{\boldsymbol{\theta_i} - \beta_q}{\boldsymbol{\alpha_q}}]$$

$\boldsymbol{\theta_i}$和$\boldsymbol{\alpha_q}$不再是单一值，而是一个矩阵。

延展2：二元到定序

Logit → Cumulative logit

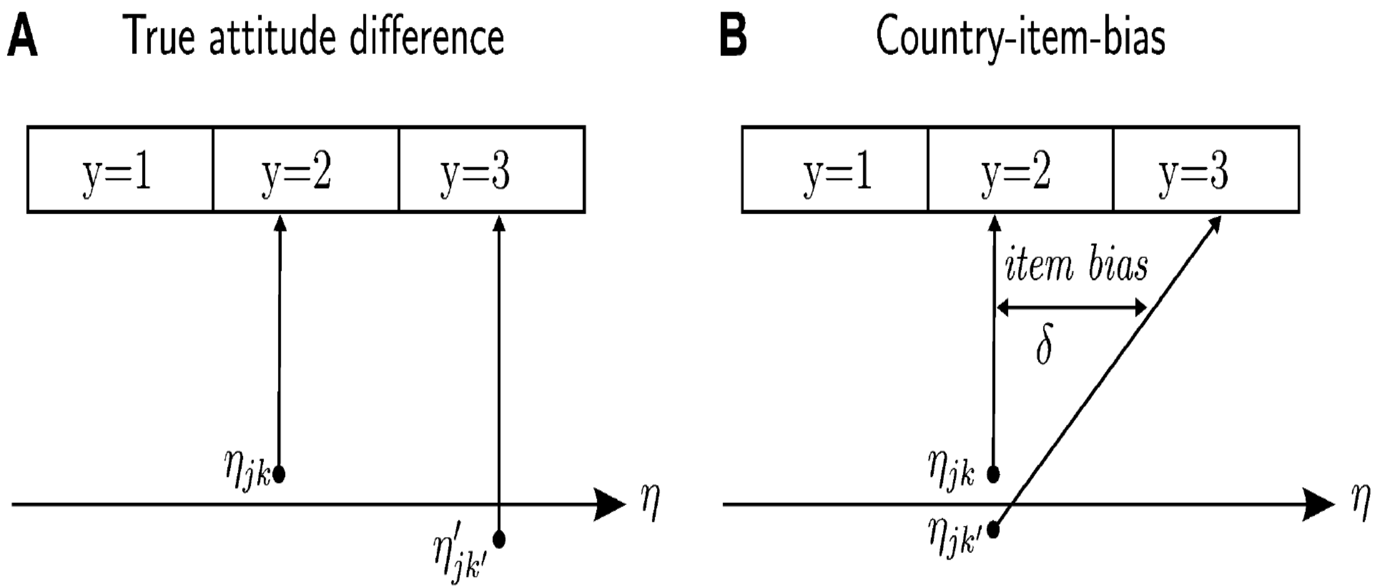
$\Pr(y_{iq} = 1) \rightarrow Pr(\frac{y_{iq} \le c}{y_{iq} > c})$



三种主要类型

1. (Modified) Graded Response Model
   - 用于scoring rubrics，比如 Likert
2. (Generalized) Partial Credit Model，Rating Scale Model
   - 用于可转化为定序的分类变量
3. Nominal Response Model
   - 用于无序分类变量

延展3：群组效应



Multilevel Mixture IRT with Item Bias Effects (Stegmueller 2011)

在估测α_q时加入random effect.

Individual fallacy: Ecological fallacy 的反面

再比如，民主、不平等、政治文化……

$$y_{kq} = \Sigma y_{ikq} / n.$$

问题：

1. 如果群组过小，其平均值的代表意义不大
2. 不同的指标对于潜在变量贡献不一样

经过群组信息（地理、人口）加权的平均值

1. 将总体（population）按群组（strata，如国家、地区）切分；
2. 估测对象为核心变量在每个群组中的平均值/比例，θ_h ($h \in \{1, H\}$);
3. 已知各群组以人口变量j（如老年男性、青年女性等）划分，群组人口（N_j）或占总人口比；
4. 各组总体平均值μ_j可通过multilevel model 进行估算。

$$\theta_h = \frac{\sum_{j \in h} N_j \mu_j}{\sum_{j \in h} n_j}$$

N: 总体（来自普查）

n: 样本（来自sample）

操作案例

数据：某年某市五区域2396家产业公司的财政信息

目标：估测每个区域的产业平均收入（记为$\theta_{1\sim5}$）

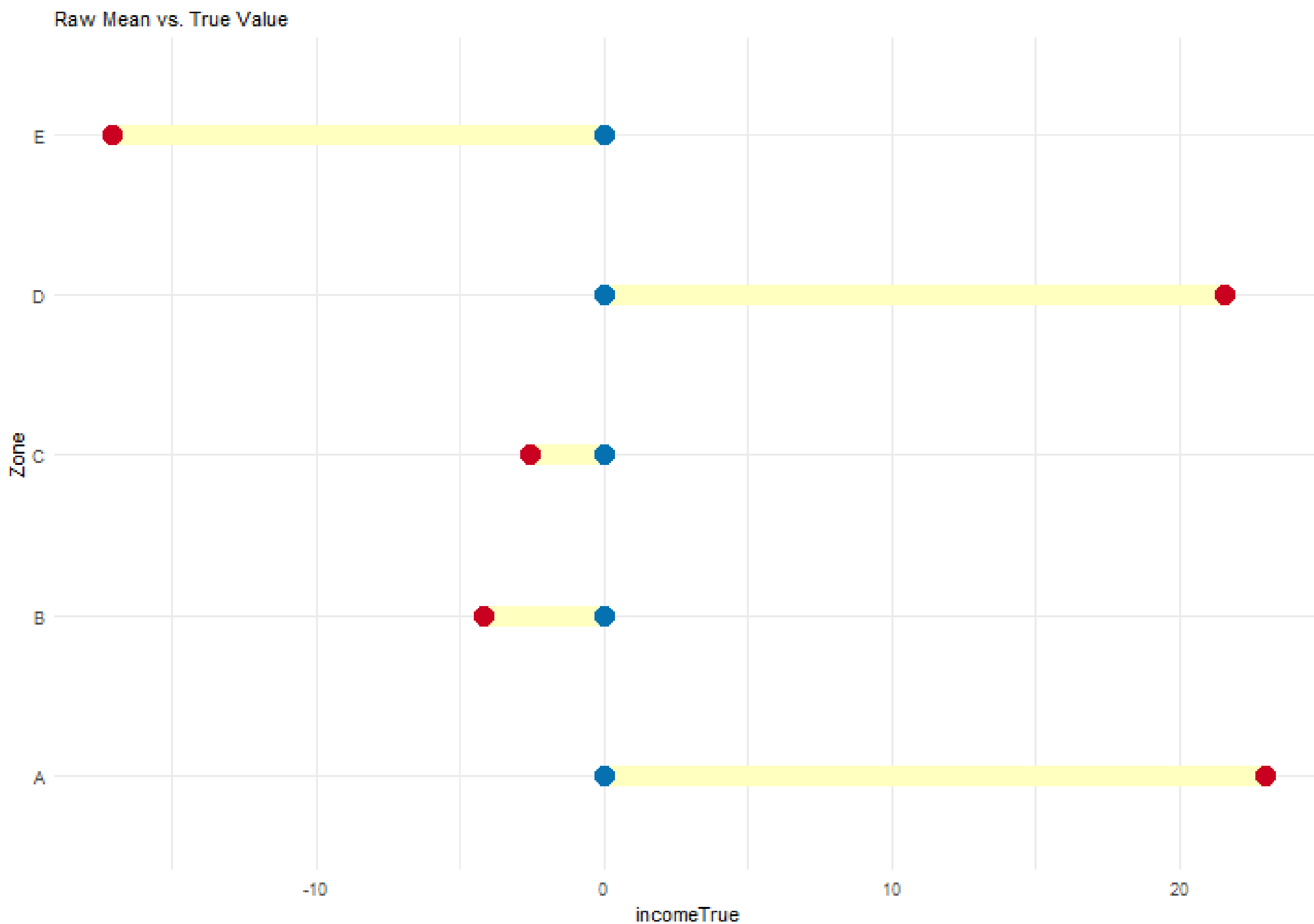
公司规模和区域分布

	A	B	C	D	E
Big	30	13	1	16	23
Medium	180	121	111	187	138
Small	97	593	862	20	4

总体平均值（真值）

Zone	income
A	652.28
B	320.75
C	331.02
D	684.98
E	767.39

我们随机选取数据中1000个产业公司作为样本：



Step I: Mr

$$Income = \beta_{0z} + \beta_{1z}Level_{iz} + \varepsilon_{iz}$$

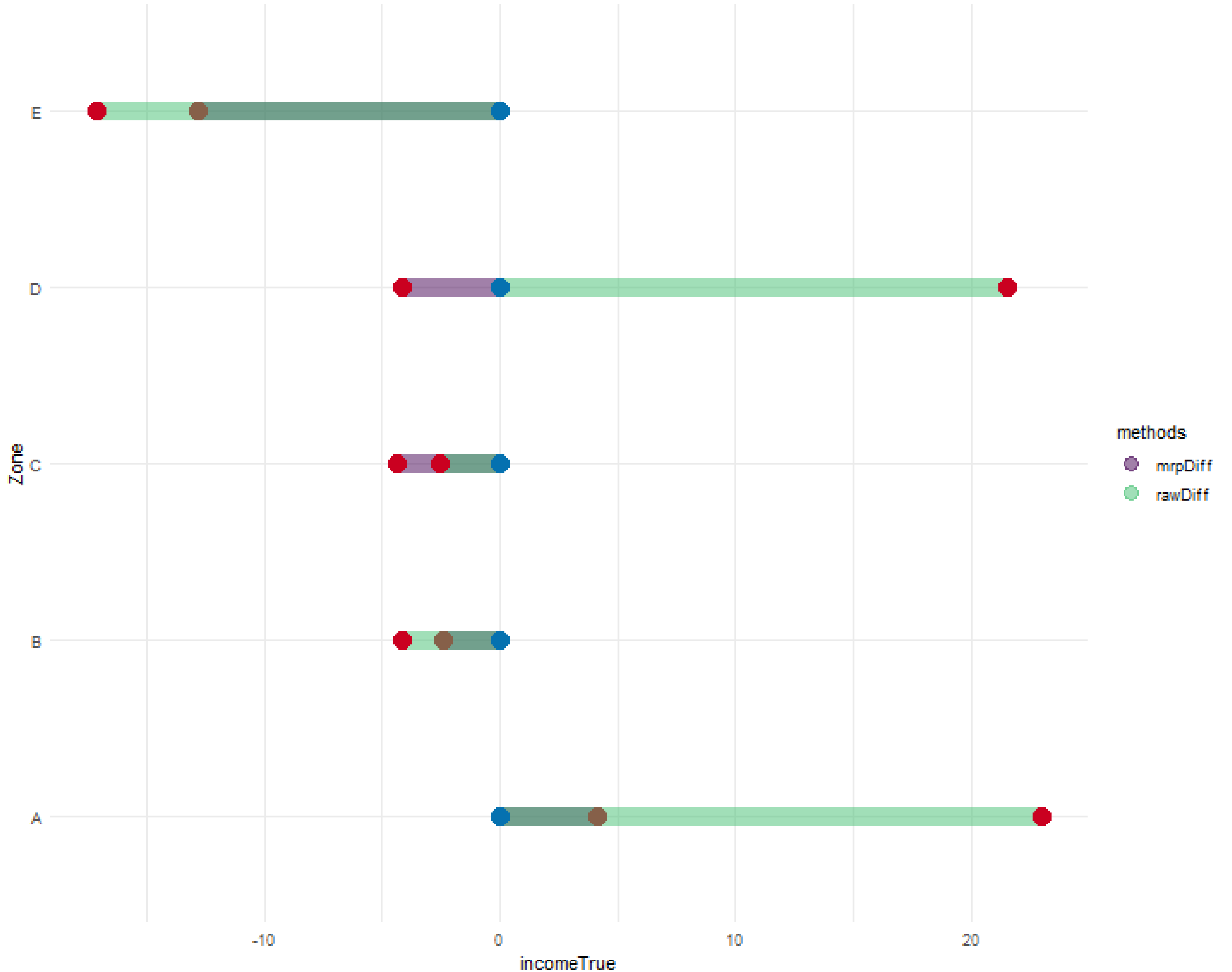
$$\beta_{0z} = \gamma_{00} + \gamma_{01}Zone_z + u_{0z}$$

	A	B	C	D	E
Big	1274.74	1148.58	1189.59	1238.51	1251.95
Medium	706.19	580.03	621.03	669.96	683.40
Small	372.95	246.79	287.79	336.72	350.16

$N_z \times$ weighted mean / n_z

```
##         A        B        C        D        E
## 656.4551 318.3753 326.6951 680.8675 754.5761
```

Comparision



- 答题难度的地区差异
- 题目的scale
- Measurement error

聚合层级IRT： DGIRT

Dynamic Group-level IRT——结合IRT和MrP (Caughey & Warshaw 2015)

DGIRT

1. 在群组层面估测IRT;
2. 在估测IRT过程中加入群组级别变量;
3. 将时间变量融入IRT估测;
4. 用MrP给估测进行权重。

IRT的群组层级估测

个体

$$p_{iq} = logist^{-1}[\frac{\theta_i - \beta_q}{\alpha_q}]$$

群组

$$\eta_{ktq} = logit^{-1}(\frac{\bar{\theta}_{kt} - \beta_q}{\sqrt{\alpha_q^2 + (1.7\sigma_{kt})^2}}).$$

$\bar{\theta}_k$ 和 σ_{kt} 是潜在变量在组k时间t的均值和sd。

囊括时间与空间问题

$$\bar{\theta}_k \sim N(\xi_t + \boldsymbol{x}'_{\boldsymbol{k}}\boldsymbol{\gamma}, \sigma^2_{\bar{\theta}})$$

$$\xi_t \sim N(\xi_{t-1};\sigma_\gamma{}^2)$$

$$\gamma_{pt} \sim N(\gamma_{p,t-1}\delta_t + z'_{p.}\eta_t, \sigma_\gamma{}^2)$$

$$n^*{}_{kqt}$$

DGIRT：

- 囊括诸多因素
- 可以部分平衡样本代表性问题
- 强大，但复杂

DGIRT简装版 (Claassen 2019)

简化1：只作用于代表性样本和国家级别

简化2：将国家作用从估测θ变为估测difficulty

简化3：忽略本地问题分布（如极化现象）

$$\eta_{ktq} = logit^{-1}(\frac{\bar{\theta}'_{kt} - \beta_q}{\sqrt{\alpha_q^2 + (1.7\sigma_{kt})^2}}).$$

$$\downarrow$$

$$\eta_{ktq} = logit^{-1}(\frac{\bar{\theta}'_{kt} - (\beta_q + \delta_{kq})}{\alpha_q}).$$

聚合IRT最新进化态：DCPO

Dynamic Comparative
Public Opinion

复杂程度：

Claasseen 2019 <
DCPO <
DGIRT

	McGann (2014)	Claassen (2019)	Caughey, O'Grady, and Warshaw (2019)	DCPO
Cross-National	✗	**YES**	**YES**	**YES**
Dynamic Priors	✗	**YES**	**YES**	**YES**
Ordinal	✗	✗	**YES**	**YES**
δ_{kq}	✗	**YES**	✗	**YES**
Bounded Mean Opinion	**YES**	✗	✗	**YES**
Opinion Polarization	**YES**	✗	✗	**YES**

优化效果

Model	Internal Validation Test (1) Mean Absolute Error (MAE)	(2) Country-Means MAE	(3) % Improvement in MAE	External Validation Test (4) k-fold Mean MAE	(5) k-fold Mean % Improvement	(6) k-fold 80% Credible Interval Coverage
Claassen (2019) Model 5	0.032	0.112	71.4	0.057	51.7	+4.9
Caughey, O'Grady, and Warshaw (2019)	0.049	0.186	73.7	0.063	66.1	-67.4
DCPO	**0.031**	**0.186**	**83.3**	**0.055**	**70.5**	**-4.5**

操作过程

1. 收集survey数据，明确与感兴趣的变量相关的指标问题
2. 通过`DCPOtools`对数据进行预处理
3. 通过`DCPO`进行数据分析
4. 通过`shinystan`诊断convergence

总结

- 个体IRT
  - Rasch model
  - nPL model
  - 多维IRT
- 检验
  - Glbal: G2/M2
  - Item: in/outfit
- 群体IRT
  - MrP
  - DGIRT
  - DCPO