



大数据背景下精准科研信息服务

2020 年版

作者：王敏杰

时间：2020-04-09

版本：0.1



Victory won't come to us unless we go to it. — M. Moore

目录

作者简介	1
1 前言	2
1.1 进度表	2
1.2 需要的配套	2
1.3 关于本文档	2
2 川师大数据	3
2.1 全景对比	3
2.2 学科对比	3
3 学科预测	4
3.1 统计方法	4
3.2 数学学科	4
3.3 物理学科	4
3.3.1 数据	4
3.3.2 先验概率	5
3.4 化学学科	6
3.5 工程学科	6
3.6 计算机学科	6
3.6.1 预测	7
3.7 其他学科	10
4 学院对学科的贡献	11
4.1 研究规模贡献分析	11
4.2 学术影响力贡献分析	11
5 选刊倾向与期刊推荐	12
5.1 各学科论文在各等级期刊上的分布	12
5.2 期刊推荐	12
A 统计口径	13
A.1 学科分类以及各学科进入 ESI 的阈值	13
A.2 数据来源	13
A.3 获取方法	13
A.4 学校列表	13
A.5 物理学科模型参数	14

A.6 参考文件	15
--------------------	----

作者简介

王敏杰，四川师范大学研究生公选课《数据科学中的 R 语言》授课老师，西南交通大学量子物理学博士，爱好数据科学，喜欢用 R 和 stan 编程，联系方式 38552109@qq.com

第 1 章 前言

1.1 进度表

- 文献调研（3月底完成）
- 数据获取（4月中旬完成）
- 数学分析和模型评估（5月中旬完成）
- 可视化（6月初完成）
- 报告初稿（6月底完成）
- 研讨会（待定）
- 正式稿发布（7月初）

1.2 需要的配套

- 需要一名学生，协助完成数据收集和整理工作（图书馆提供劳务费）

1.3 关于本文档

本报告使用 R 和 stan 语言完成，数据和代码存放在 GitHub 仓库<https://github.com/perlatex/ElegantBookdown4IS>，欢迎批评指正。

第 2 章 川师大数据

2.1 全景对比

横向比较 top30 所师范类高校的学科发展情况

- 全景大图高亮的（包括川师在内的四个学校）类似 R4DS 中的 eda_covid2019 吸取 Kieran Healy 大神的配色方案
- 学科小图

2.2 学科对比

分面各校，高亮川师

第3章 学科预测

本章的主要工作是，计算并预测川师未来三年进入双一流学科的概率。可能一点意义也没有

3.1 统计方法

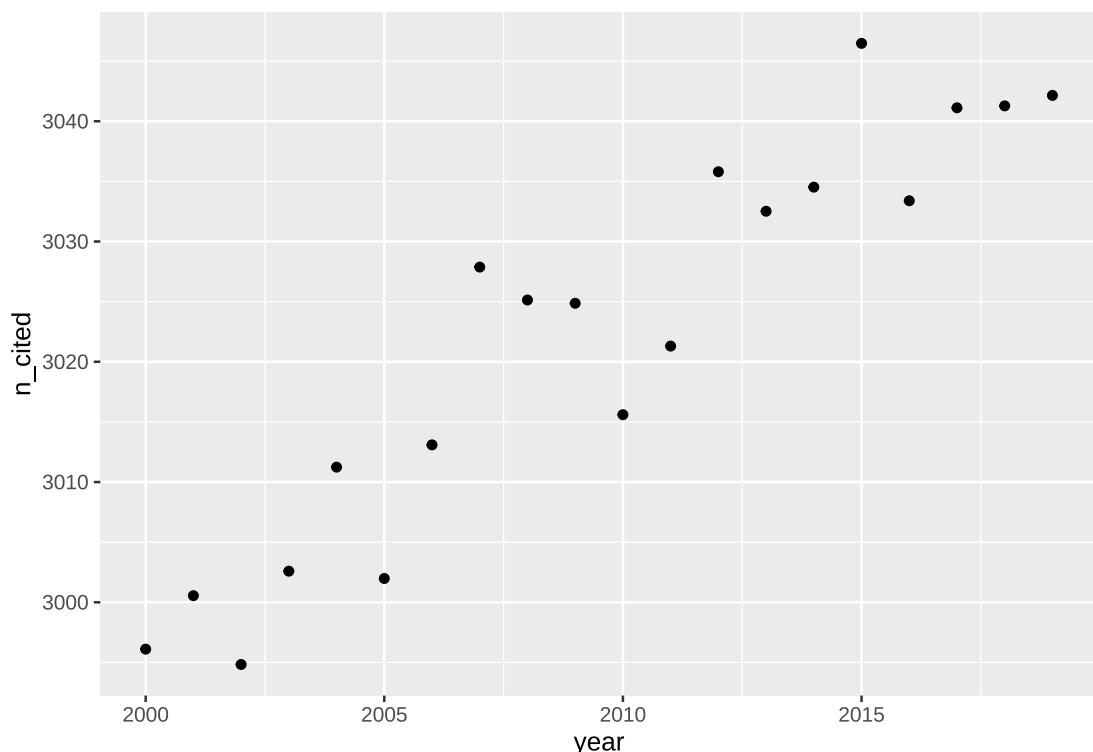
- 贝叶斯数据分析
- 模型不好怎么办？
- \log_{10} scale

关于模型，学科的发展和很多方面都有关系，因此建立一个完全正确的模型是不可能的。正如英国统计学家 George E. P. Box 所说，所有模型都是错的，但其中有些是有用的。所以与其去建立复杂的模型，并给解释带来更多困扰，不如就从最简单的出发。

3.2 数学学科

3.3 物理学科

3.3.1 数据



3.3.2 先验概率

```

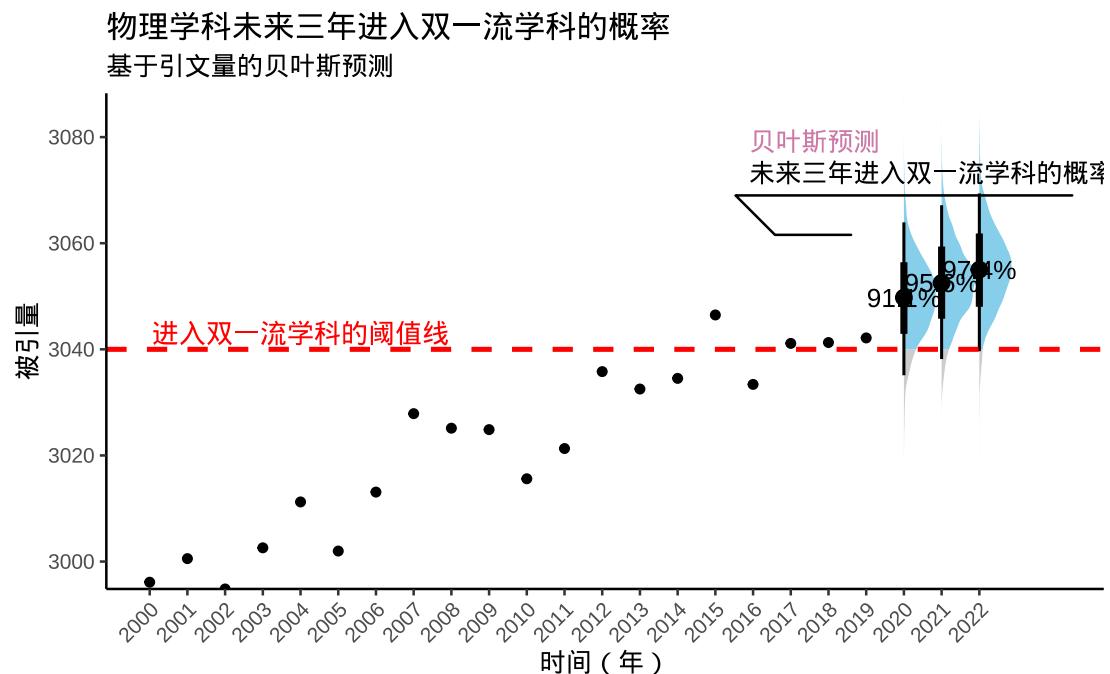
#> Family: gaussian
#> Links: mu = identity; sigma = identity
#> Formula: n_cited ~ 1 + year
#> Data: d2 (Number of observations: 20)
#> Samples: 4 chains, each with iter = 41000; warmup = 40000; thin = 1;
#>          total post-warmup samples = 4000
#>
#> Population-Level Effects:
#>             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
#> Intercept -2266.03     516.07 -3265.95 -1229.33 1.00      2670     2280
#> year        2.63      0.26     2.12     3.13 1.00      2671     2294
#>
#> Family Specific Parameters:
#>             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
#> sigma       6.51      1.20     4.68     9.34 1.00      2543     2096
#>
#> Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
#> and Tail_ESS are effective sample size measures, and Rhat is the potential
#> scale reduction factor on split chains (at convergence, Rhat = 1).

#> # A tibble: 12,000 x 6
#> # Groups:   year, .row [3]
#>   year   .row .chain .iteration .draw .prediction
#>   <dbl> <int> <int>     <int> <int>     <dbl>
#> 1 2020     1    NA       NA     1 3049.
#> 2 2020     1    NA       NA     2 3052.
#> 3 2020     1    NA       NA     3 3064.
#> 4 2020     1    NA       NA     4 3056.
#> 5 2020     1    NA       NA     5 3044.
#> 6 2020     1    NA       NA     6 3044.
#> 7 2020     1    NA       NA     7 3051.
#> 8 2020     1    NA       NA     8 3056.
#> 9 2020     1    NA       NA     9 3041.
#> 10 2020    1    NA       NA    10 3056.
#> # ... with 11,990 more rows

#> # A tibble: 3 x 3
#>   year pred_mean prob_above_line

```

```
#>      <dbl>      <dbl>      <dbl>
#> 1    2020     3050.     0.911
#> 2    2021     3053.     0.956
#> 3    2022     3055.     0.974
```



3.4 化学学科

3.5 工程学科

3.6 计算机学科

具体参考 https://mc-stan.org/docs/2_22/stan-users-guide/prediction-forecasting-and-backcasting.html

```
#> Inference for Stan model: simple.
#> 4 chains, each with iter=41000; warmup=40000; thin=1;
#> post-warmup draws per chain=1000, total post-warmup draws=4000.
#>
#>           mean   se_mean     sd    2.5%    25%    50%    75%   97.5%
#> alpha     -2007.74     3.83 102.57 -2211.51 -2074.88 -2008.05 -1941.31 -1800.92
#> beta       2.50     0.00   0.05    2.40    2.47    2.50    2.54    2.60
#> sigma      6.35     0.03   1.07    4.65    5.59    6.19    6.97    8.79
#> new_y[1]  3048.54     0.11   6.59 3035.59 3044.20 3048.30 3052.82 3061.71
#> new_y[2]  3051.05     0.11   6.74 3037.94 3046.68 3051.01 3055.41 3064.60
```

```

#> new_y[3] 3053.55    0.10   6.63   3040.36   3049.32   3053.61   3058.00   3066.21
#> lp_-      -63.35    0.05   1.26   -66.57   -63.94   -63.03   -62.44   -61.94
#>           n_eff Rhat
#> alpha      718 1.01
#> beta       717 1.01
#> sigma      1148 1.00
#> new_y[1]  3784 1.00
#> new_y[2]  3863 1.00
#> new_y[3]  4034 1.00
#> lp_-      755 1.00
#>
#> Samples were drawn using NUTS(diag_e) at Thu Apr 09 18:56:26 2020.
#> For each parameter, n_eff is a crude measure of effective sample size,
#> and Rhat is the potential scale reduction factor on split chains (at
#> convergence, Rhat=1).

```

3.6.1 预测

```

#> # A tibble: 12,000 x 5
#> # Groups:   condition [3]
#>   condition new_y .chain .iteration .draw
#>   <int> <dbl> <int>     <int> <int>
#> 1     1 3048.     1         1     1
#> 2     1 3047.     1         2     2
#> 3     1 3049.     1         3     3
#> 4     1 3044.     1         4     4
#> 5     1 3058.     1         5     5
#> 6     1 3041.     1         6     6
#> 7     1 3054.     1         7     7
#> 8     1 3051.     1         8     8
#> 9     1 3050.     1         9     9
#> 10    1 3050.     1        10    10
#> # ... with 11,990 more rows

#> # A tibble: 12,000 x 6
#>   condition new_y .chain .iteration .draw year
#>   <int> <dbl> <int>     <int> <int> <dbl>
#> 1     1 3048.     1         1     1 2020
#> 2     1 3047.     1         2     2 2020
#> 3     1 3049.     1         3     3 2020

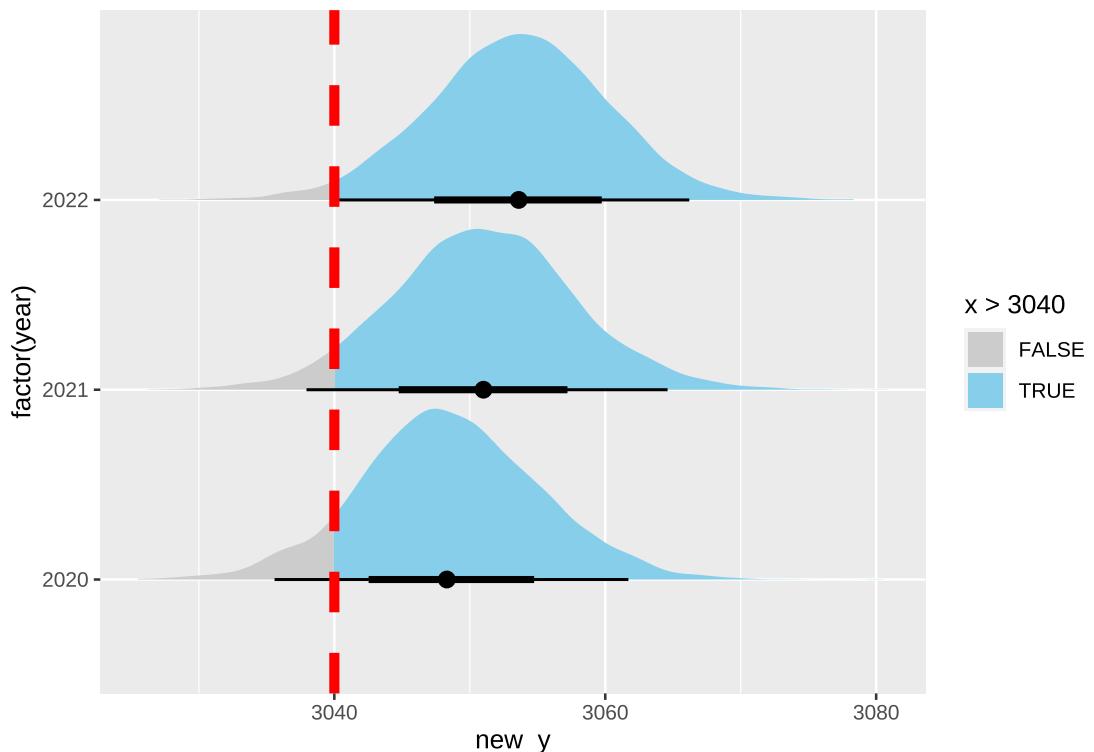
```

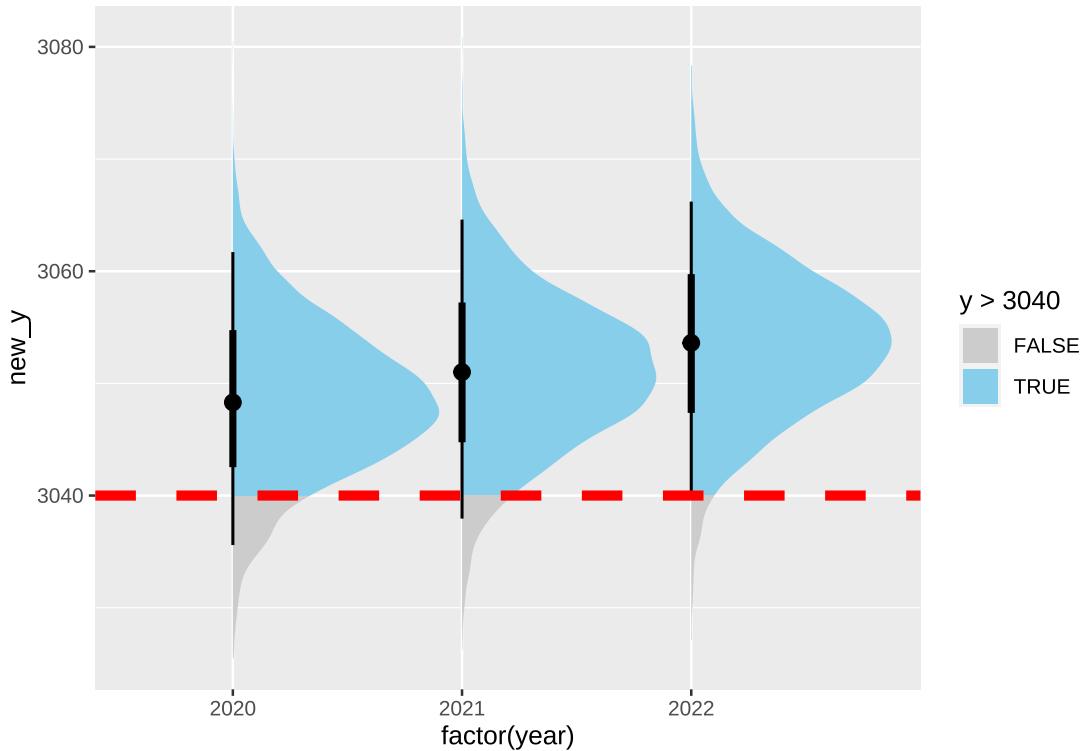
```

#> 4      1 3044.      1      4      4 2020
#> 5      1 3058.      1      5      5 2020
#> 6      1 3041.      1      6      6 2020
#> 7      1 3054.      1      7      7 2020
#> 8      1 3051.      1      8      8 2020
#> 9      1 3050.      1      9      9 2020
#> 10     1 3050.      1     10     10 2020
#> # ... with 11,990 more rows

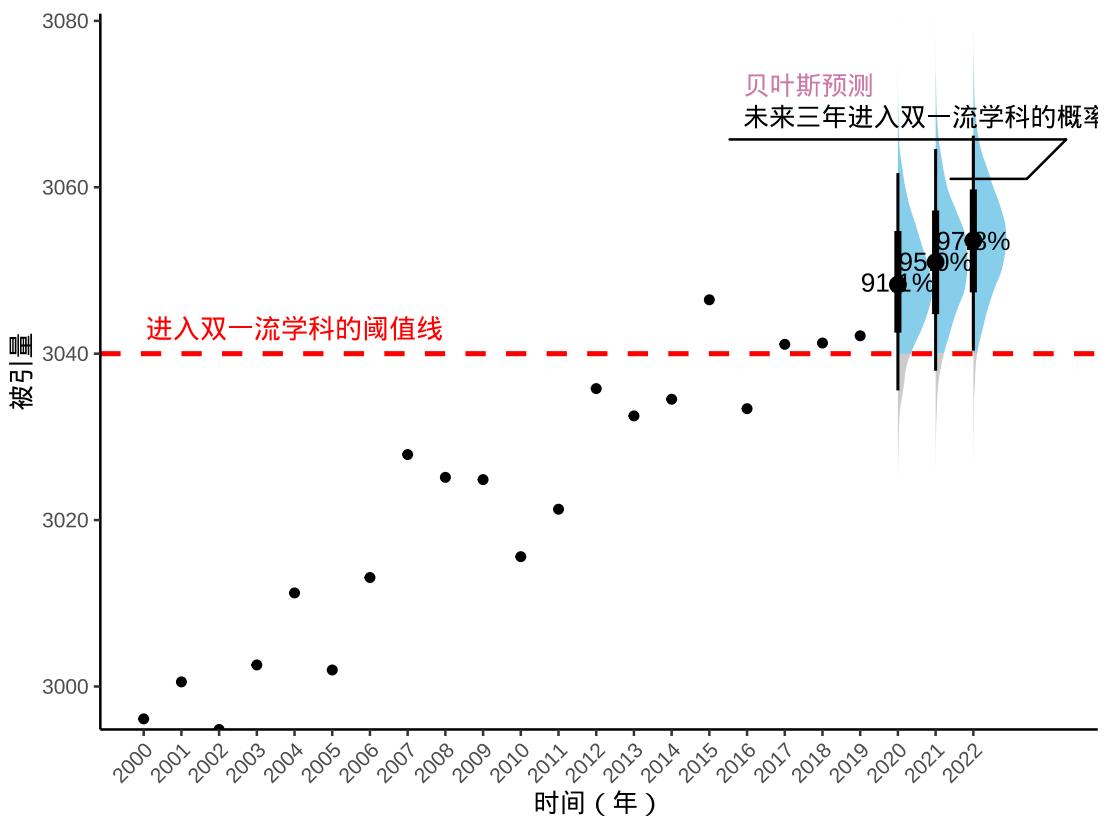
#> # A tibble: 3 x 3
#>   year pred_mean prob_above_line
#>   <dbl>     <dbl>          <dbl>
#> 1 2020     3049.        0.911
#> 2 2021     3051.        0.950
#> 3 2022     3054.        0.978

```





物理学科未来三年进入双一流学科的概率
基于引文量的贝叶斯预测



3.7 其他学科

如果需要了解其他学科的信息，请联系本文作者¹

¹38552109@qq.com

第4章 学院对学科的贡献

4.1 研究规模贡献分析

数据不准确，避免引起歧义。暂时不开展

4.2 学术影响力贡献分析

数据不准确，避免引起歧义。暂时不开展

第5章 选刊倾向与期刊推荐

5.1 各学科论文在各等级期刊上的分布

数据不准确，避免引起歧义。暂时不开展

5.2 期刊推荐

数据不准确，避免引起歧义。暂时不开展

附录 A 统计口径

A.1 学科分类以及各学科进入 ESI 的阈值

ESI 学科分类一种较为宽泛的学科分类模式。ESI 学科分类模式基于期刊分类，由自然科学与社会科学的 22 个学科构成。艺术与人文期刊没有被包含。每一本期刊只被划分至 22 个 ESI 学科中的一个，没有重叠的学科设置使得分析变得更为简单。被归类为跨学科学科 (Multidisciplinary field) 的 Science、Nature 与 PNAS 期刊，会被按照各篇文章的参考文献 (reference) 与引用文献 (citation)，重新为每篇文章单独分类，但每篇文章仍只会被分类到一个学科。

A.2 数据来源

- 用 ESI 不用 wos
- 2010 - 2019 十年，6 个学科（数学，物理，化学，工程，计算机）
- 获取下载地址
 - 链接 1，检索学校历年发文量的 (<https://incites.clarivate.com/zh/#/explore/0/subject>)
 - 链接 2，近期进入 ESI 学科的阈值 (<https://esi.clarivate.com/ThresholdsAction.action>)

A.3 获取方法

整理的 raw-data 可以在这里找到 <https://github.com/perlateX/ElegantBookdown4IS/tree/master/data>

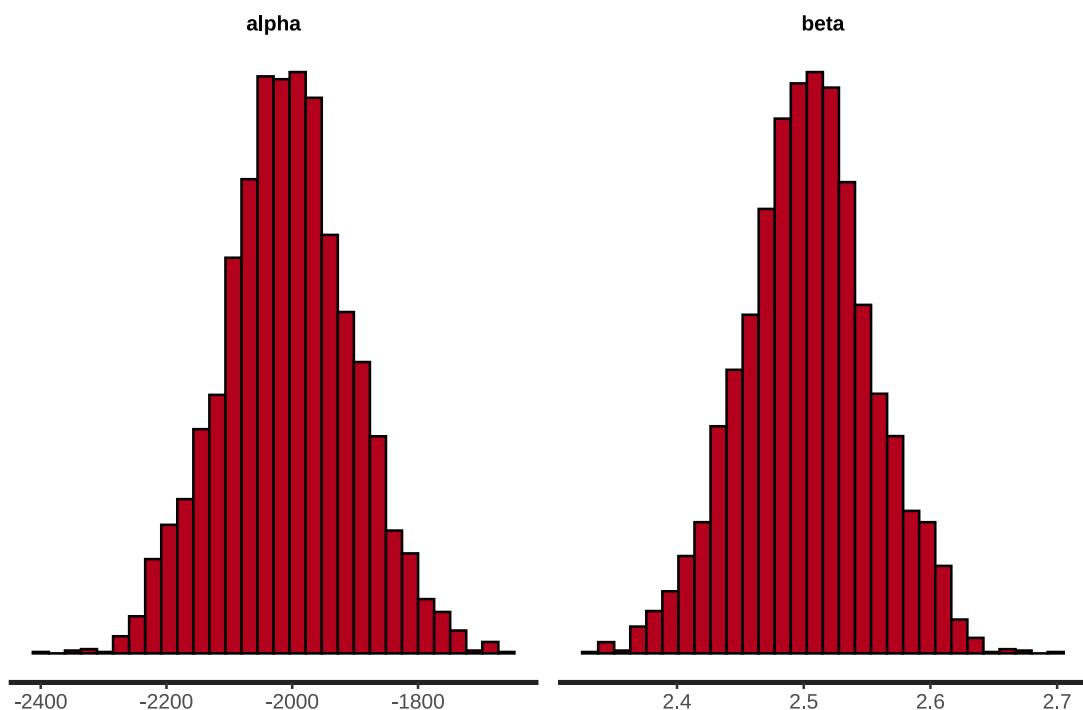
A.4 学校列表

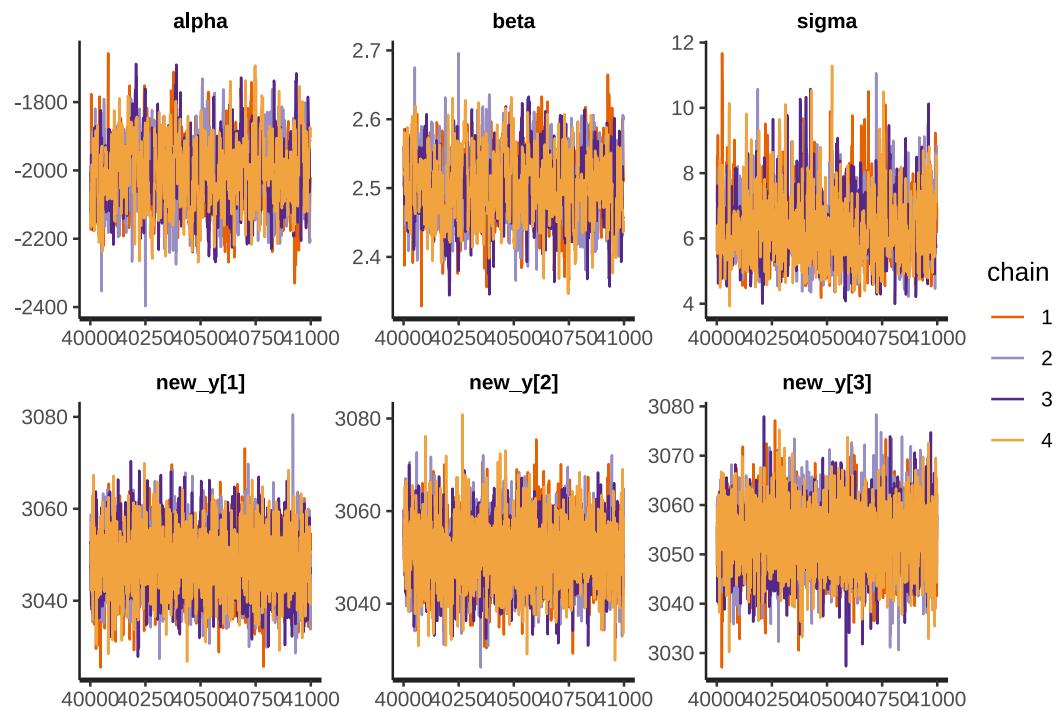
- 师范类学校清单 (<https://www.dxsbb.com/news/1448.html>)
- 选取依据 (top30) (川师 25 名)
- 省份中文名英文名

表 A.1: ESI 学科分类以及各学科进入 ESI 的阈值 (2020 年 3 月数据)

category	discipline	threshold202003
工学 (3)	计算机科学 (Computer Science)	1692
	工程科学 (Engineering)	5079
	材料科学 (Materials Sciences)	5981
生命科学 (4)	生物与生化 (Biology & Biochemistry)	1855
	环境 / 生态学 (Environment/Ecology)	2837
	微生物学 (Microbiology)	3549
	分子生物与遗传学 (Molecular Biology & Genetics)	1876
社会科学 (2)	一般社会科学 (Social Sciences, General)	3319
	经济与商学 (Economics & Business)	4795
理学 (5)	化学 (Chemistry)	3844
	地球科学 (Geosciences)	3918
	数学 (Mathematics)	3620
	物理学 (Physics)	4421
农学 (2)	空间科学 (Space Science)	10243
	农业科学 (Agricultural Sciences)	2087
	植物与动物科学 (Plant & Animal Science)	4959
医学 (5)	临床医学 (Clinical Medicine)	2864
	免疫学 (Immunology)	14029
	神经科学与行为 (Neuroscience & Behavior)	2236
	药理学与毒物学 (Pharmacology & Toxicology)	3464
其他 (1)	精神病学 / 心理学 (Psychology/Psychiatry)	1142
	多学科 (Multidisciplinary)	27851

A.5 物理学科模型参数





A.6 参考文件

- stan
- ggplot
- tidybayes
- tidyverse
- 书籍