STAT243 Lecture 7.1 A Few Preparatory Notes

1 关于“大数据”的思考 (An Editorial on “Big Data”)

- **热度的变化**
  - “大数据 (big data)”曾经是热门词汇，但随着 AI/ML 革命的到来，它不再是主要的流行语。
  - 然而，AI/ML 的发展本身正是建立在**大规模可获取数据集**的基础上。
- **理性看待“大数据”**
  - 部分“大数据”热潮是合理的，因为：
    - 大规模数据能支持更复杂、更非线性的模型
    - 能回答传统小样本无法探索的问题
  - 但“大数据”并不能解决以下根本问题：
    - **相关不代表因果 (correlation ≠ causation)**
      例如：即使拥有全美人口的健康数据，也无法仅凭观测得出“盐摄入导致高血压”
    - **非随机样本 ≠ 统计样本**
      非代表性样本无法推断总体特征
      如 COVID-19 数据虽多，但因不完整/不均衡而难以得出可靠结论
    - **多重检验 (multiple testing)**
      在庞大数据上执行大量分析会提高“伪显著”概率，纯属噪声的结果可能看起来“有意义”
- **结论**
  - 一个**小而精的代表性数据集**往往比一个庞大但杂乱的样本更有信息价值
  - 大数据仍有价值：可用于筛选、构造更合理的样本，或支持因果推断设计
  - 但必须搭配严谨的统计思考与实验设计使用
- **“大”的不同定义**
  - **技术定义**：指数据体量或采集速度极大
  - **社会定义**：指数据与实证分析在社会中的普遍应用，而非数据量本身

> 延伸阅读：
> 文章讨论电子健康记录 (EHR) 分析中的偏差问题
>
> - https://doi.org/10.1093/jrsssa/qnae039
> - https://doi.org/10.1093/jrsssa/qnae005

---

2 数据规模与运算限制 (Logistics and Data Size)

- **内存瓶颈**
  - Python 与 R 的主要限制：所有对象都存储在内存中
  - 因此单机处理的数据集通常受限于 **1–20 GB**，具体取决于系统内存
- **适用范围**
  - 本单元介绍的技术主要针对：
    - **GB 到数十 GB** 的数据规模
    - 若机器拥有足够内存或磁盘，也可扩展至更大规模
  - 若数据规模达到：
    - **数百 GB – TB 级别** → 需使用数据库或云平台
    - **PB 级别** → 建议采用 Spark、AWS、GCP 等大数据框架
- **数据存储建议**
  - 将大文件存放于**本地磁盘**而非网络驱动器
  - 避免网络传输带来的延迟与带宽瓶颈

3 我们已掌握的“大数据”处理基础 (What We Already Know About Handling Big Data)

- **UNIX 命令的高效性**
  - UNIX 命令行操作速度极快，适合大文件的行/列筛选
  - 可使用以下命令组合实现高效数据预处理：
    - `grep`：按条件提取行
    - `head` / `tail`：查看文件头尾
    - `awk`：按条件提取行或列
  - 结合 **管道操作 (piping)** 与 **shell 脚本** 可快速缩减数据规模
- **GNU Parallel**
  - 命令行并行化工具
  - 可在 Linux 集群中同时执行多个任务，提高处理效率
- **实践建议**
  - **删除无关列**：若数据有 30 列而仅需 5 列，应先提取子集
  - **抽样分析**：在许多情况下，用随机样本的结果与全量分析几乎一致
  - 通过这些简化策略，可继续使用熟悉的 Python/R 工具，而无需复杂的大数据框架
- **存储格式优化**
  - 二进制格式 (如 `.parquet`、`.feather`) 比文本格式 (如 `.csv`) 更紧凑、高效
  - 推荐在分析前将大型文本数据转换为二进制格式
- **使用数据库**
  - 对许多应用而言，将大型数据集导入 **标准数据库系统** (如 PostgreSQL、SQLite、DuckDB) 是一种稳健且高效的解决方案