回归分析学习指导

23 秋《回归分析》课程参考

组织：中国科学技术大学

时间：January 6, 2024

版本：4.3

编者：助教

世界上的事情，最忌讳的就是十全十美，你看那天上的月亮，一旦圆满了，马上就要亏厌，树上的果子，一旦熟透了，马上就要坠落。但是总要稍留欠缺，才能持恒。——莫言

目录

第 1 章 课程介绍 **1**

第 2 章 预修内容 **2**

2.1 概率论 2

2.2 数理统计 2

2.2.1 抽样分布、预备知识 2

2.2.2 点估计 3

2.2.3 区间估计 4

2.2.4 参数假设检验 4

2.2.5 一些非参数假设检验方法 6

2.3 R 语言编程 6

2.4 线性代数 6

2.4.1 方阵的谱分解 7

2.4.2 矩阵的奇异值分解 7

2.4.3 广义逆 7

2.4.4 欧式空间中的正交投影 7

第 3 章 似然方法及其理论 **8**

3.0 Kullack-Leibler 距离 8

3.1 极大似然估计 8

3.2 极大似然估计的渐近性质 10

3.3 基于似然的检验：简单原假设情形 12

3.4 基于似然的检验：复合原假设情形 12

第 4 章 作业解答 **15**

4.1 HW 1 15

4.2 HW 2 16

4.3 HW 3 19

4.3.1 习题 19

4.3.2 补充内容 21

4.4 HW 4 22

4.5 HW 5 25

4.6 HW 6 28

4.7 HW 7 31

4.8 HW 8 36

4.9 HW 9 40

4.10 HW 10 43

4.11 HW 11 46

4.12 HW 12 51

第 5 章 复习 **56**

5.1 因果推断 56

5.2 相关性 56

5.2.1 相关系数和偏相关系数 56
5.2.2 相关性检验 56
5.2.3 辛普森悖论 56
5.2.4 随机向量 56
5.3 线性回归模型 57
5.3.1 总体线性模型 57
5.3.2 单变量简单线性模型 57
5.3.3 多元回归 57
5.4 方差分析 58
5.4.1 单因素方差分析 58
5.4.2 两因素方差分析 58
5.5 回归诊断 58
5.5.1 残差分析 58
5.5.2 影响分析 58
5.5.3 6 个图 58
5.6 预测 58
5.6.1 预测误差分解 58
5.6.2 线性模型的预测 59
5.6.3 变量选择 59
5.6.4 最优子集回归 59
5.6.5 变量选择的贪心算法 59
5.7 写在最后 59

第 1 章 课程介绍

《回归分析》，是一门研究线性模型的学科，同学们高中时学到的“最小二乘法”“回归直线”等概念便是这门课的核心。这门课用到的工具是概率统计方法与线性代数，并没有真正数学意义上的“分析”。我们通过建立线性模型来说明变量之间的线性相关关系。

线性模型，也是当下最热门的发展方向“机器学习”里最简单的一个模型，但却有许许多多有趣的东西可以讲。

可以说，如果你去学习《机器学习》，那么会学到各种各样的模型：决策树、支持向量机、神经网络、聚类、集成……它们有的带参数、有的不带；有的带响应变量，也就是

$$y_i = kx_i + b + \varepsilon_i$$

里面的响应变量 y_i，我们称之为“有监督学习”，线性模型便是其中最简单的一种，当然还有相对应的“无监督学习”，我们只知道各个样本的特征信息，并不知道它的类别信息。但几乎所有的“机器学习”课程为了介绍得更多、更全面，对于每一个具体的模型、或者叫方法，都是浅尝辄止，在介绍完基本的模型以及算法以后，相关理论一带而过。这也是因为当下对于机器学习的实验研究发展速度过快，统计理论无法紧随其后充分解释的结果。

但如果你学《回归分析》，整个一学期都是在研究最简单的“线性模型”——在一定假设条件下，没有什么理论本质上的困难。我们从统计的角度充分去探索线性模型的有趣之处，并且从各个角度来解释我们得到的结果，学下来以后会有整体思路十分清晰、想得到的问题都能被解决的愉快感觉。当然，这里会为你打下坚实的理论基础，也算是对统计与机器学习理论的初窥。

课程内容的简单介绍：

- 因果论断：启发学这门课的原因，关联和因果的关系是什么？我们用什么样的方法可以得到怎样的结论？在学完这门回归课以后，用线性模型可以做到什么？
- 相关性分析：随机变量的相关性，两个随机向量整体的相关性（用一个数来表示两个随机向量的相关性、而不是直觉上想到的协方差、相关系数矩阵），控制变量以后的相关性即偏相关。这部分学起来似乎和回归毫无关联，计算也略微繁琐，但后面发现这其实就是回归模型的起源，一脉相承。
- 线性模型的多视角阐述，本课程的重头戏，从最简单能直观感受到的一元简单线性模型到多元的最小二乘方法，你几乎不可能单从任何一本书甚至几本书上学到如此完整的线性模型理论。
- 方差分析：判断不同效应的影响是否相同的一套假设检验理论，可以直接写成线性模型 F 检验的形式。
- 回归诊断：如何说明回归是否有效？
- 预测：很快带过，这部分其实是机器学习的重头戏。

当然，对大部分同学来说这门课还是相当容易的，如果你愿意上课认真学习的话。而且杨老师的做的课件可以说是科大统计课的标杆，内容丰富充实，既有与课程相关的全部理论推导，又有许多老师自己的理解加上一些前沿问题的介绍。所以我甚至觉得我没有太多可以写的东西，希望这个讲义在“预修内容”以后不会烂尾。当然，如果我后面什么也没有再补充了，那也是对老师的极大肯定。

第 2 章 预修内容

我是在 2022 年秋季学期选修的这门课，回归分析是我统计的入门课之一，也就是说，我并未在之前学过有关统计的内容。

如果是管理学院统计系的同学，概率论与数理统计都已经相当熟悉，那么这部分内容基本可以直接跳过，或者简单看一下回归分析里用得较多的一部分线性代数。当然，根据往年惯例，老师会将课上用到的同学们可能会不熟悉的线性代数内容重新讲解一遍，所以完全不必担心基础问题。

对于数学学院概率统计或者计算方向的同学，直接在大三上学期与数理统计同时选修这门课还是有一定的压力，并且学校甚至将这两门课安排成了时间冲突。如果不提前自学一些相应内容，或者在学期内的学习过程中没有比较好的“缺什么补什么”的自主学习能力，可能在上这门课时会面临比较大的压力，所以往年可能只有少部分同学顶着一些压力选管院的这门《回归分析》。但问题在于，数院概统同学在大三的秋季学期可以选择的课程太少了，不应该错过这门好课。所以为了降低这门课的门槛，让更多的同学敢选，并且学好这门课，我会较为详细的介绍需要掌握的预修知识。不过由于时间和精力有限，有些内容也只能做到点到为止，我会尽量提供全面的参考资料，也欢迎在学习过程中产生困惑的同学多与助教交流。

其他课程其实也类似。**你可能知道选修任何统计课都要数理统计和 R 语言作基础，但你不知道的都是你只需要会一点点就足够了**。

2.1 概率论

概率论是同学们都学过的内容，这门课用到的也基本只是各种各样的概率分布形式，并不过多涉及概率论里比较深刻的定理。如果一定要说一些，那么你只要记住大数定律和独立同分布情况下的中心极限定理，应该就够了。有时间的话我会参照往年课程内容补充一些具体的例子。

2.2 数理统计

数理统计里的基本概念是学这门课之前必须要了解的。你需要了解统计学家如何思考问题，如何做推断，如何有理由地说服别人相信你的想法，或者有理有据地否定你认为错误的论断。但正如前面所说，了解最简单的一部分就足够了。

2.2.1 抽样分布、预备知识

了解数理统计的若干基本概念：总体、个体、样本、样本容量、抽样。统计中经常会将样本视作与独立同分布的随机变量 $X_1, \ldots, X_n$，与总体 X 也同分布。这样的样本称为简单随机样本。

样本分布族可能为参数空间：

$$\mathcal{F} = \{f(x, \lambda) : \lambda > 0\}.$$

其中 λ 为参数，它的每一个可能的值对应于一个具体分布。

注 样本具有两重性。既可以看作抽样时得到的具体的数，也可以看作随机变量。

定义 2.1 (统计量)

由样本算出来的量称为统计量。或者说，统计量是样本的函数。 ♣

注 统计量只与样本有关，不能与未知参数有关。

例题 2.1 设 $X_1, \ldots, X_n$ 是从某总体 X 中抽取的样本，则称

$$\bar{X} = \frac{1}{n} = \sum_{i=1}^{n} X_i$$

为样本均值。称

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

为样本方差。它反映了总体方差的信息。这里除以 $n-1$ 是为了得到总体方差 $Var(X)$ 的无偏估计，即 $\mathbb{E}S^2 = Var(X)$.

定理 2.1 (正态变量样本均值和样本方差的分布)

设 $X_1, \ldots, X_n \sim N(a, \sigma^2)$, 样本均值和方差定义如上，则有

$$\bar{X} \sim N(a, \sigma^2/n)$$

$$(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$$

♡

关于方差的卡方分布证明方法不少，应该在概率论里讲过，希望同学熟悉所有有关正态分布的命题。

定义 2.2 (三大分布)

设 $X_1, \ldots, X_n i.i.d. \sim N(0,1)$, 则称 $\xi = \sum_{i=1}^{n} X_i^2$ 是自由度为 n 的 χ^2 变量，记为 $\xi \sim \chi^2_n$.

设随机变量 $X \sim N(0,1), Y \sim \chi^2_n$ 且 X 和 Y 独立，则称 $T = \frac{X}{\sqrt{Y/n}}$ 是自由度为 n 的 t 变量，其分布称为自由度为 n 的 t 分布，记为 $T \sim t_n$.

设 $X \sim \chi^2_m, Y \sim \chi^2_n$ 且 X 和 Y 独立，则称 $F = \frac{X/m}{Y/n}$ 为自由度为 m 和 n 的 F 变量，其分布称为自由度是 m 和 n 的 F 分布，记为 $F \sim F_{m,n}$.

♣

关于三大分布，你只需要记住它们的定义（独立性是必要的）和密度曲线的形状，不需要记住它们的密度函数的具体形式。我们会在后续课程中常常构造符合这样分布的统计量来做推断。

2.2.2 点估计

点估计是利用样本产生的统计量来估计总体分布中的参数，例如我们可以用 $\bar{X}$ 来作为 $\mathbb{E}X$ 的点估计，并且由于 $\mathbb{E}\bar{X} = \mathbb{E}X$, 这样的估计被称作是无偏的。这种利用样本的矩来产生点估计的方法，称作矩估计。

另外一种常见的估计是极大似然估计。极大似然估计可以常常被证明出许多好的性质，操作起来也只是将问题转化成了一个函数的优化（求极值）问题。因此在统计等领域应用颇多。我想较为详细地介绍它的定义。

设有一参数分布族 $\mathcal{F} = \{f(x, \theta), \theta \in \Theta\}$, 其中 Θ 为参数空间。令 $X = (X_1, \ldots, X_n)$ 为从 $\mathcal{F}$ 中抽取的简单随机样本，$f(\boldsymbol{x}, \theta) = f(x_1, \ldots, x_n, \theta)$ 为样本 X 的概率（密度）函数。

定义 2.3 (似然函数)

设 $f(\boldsymbol{x}, \theta) = f(x_1, \ldots, x_n, \theta)$ 为样本 X 的概率（密度）函数。当 $\boldsymbol{x}$ 固定时，可以看成 θ 的函数，称为似然函数。

♣

注 对似然函数取对数得到的函数称为对数似然函数。它和似然函数增减性相同，且求对数可以将乘积转化为求和更容易研究，因此实际操作中经常选择极大化对数似然函数。

定义 2.4 (极大似然估计)

令 $X = (X_1, \ldots, X_n)$ 为从 $\mathcal{F}$ 中抽取的简单随机样本（独立同分布），称

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, f(\boldsymbol{x}, \theta)$$

为极大似然估计。 ♣

注 由于分布族已知，得到的 $\hat{\theta}$ 仅仅是样本 $X=(X_1,\ldots,X_n)$ 的函数，是一个统计量。

注 直观地看，我们选取 θ 使得当前事件（即样本 $X=(X_1,\ldots,X_n)$ 被产生出来）的概率最大。这一点不难理解，例如我们抽取 100 个样本发现均接近 0，我们更倾向于认为样本的均值接近于 0，而不是 1000，因为当样本均值为 1000 时产生这样的样本的可能性太小了。就是这样直观的想法，可以为我们提供如此优良的极大似然估计。在数理统计课上你会被迫做许多练习，但我认为知道最简单的定义并理解背后的想法就足够了。

2.2.3 区间估计

区间估计通常伴随假设检验产生。这里你也只需要了解最简单的定义。现在你拥有的，依然是假设出来的参数分布族 $\mathcal{F}=\{f(x,\theta),\theta\in\Theta\}$，并且你收集了 n 个简单随机样本，想要对参数 θ 的取值作出估计。你构造了一个统计量 $\hat{g}_1(\boldsymbol{X})$ 并且认为它大概率比 θ 小，另一个统计量 $\hat{g}_2(\boldsymbol{X})$ 并且认为它大概率比 θ 大，因此你有充分把握认为 θ 会落在 $\hat{g}_1(\boldsymbol{X})$ 和 $\hat{g}_2(\boldsymbol{X})$ 里。

定义 2.5 (置信区间)

如果你可以证明，$\mathbb{P}(\hat{g}_1(\boldsymbol{X})\le\theta\le\hat{g}_2(\boldsymbol{X}))\ge 1-\alpha$, 那么区间 $[\hat{g}_1(\boldsymbol{X}),\hat{g}_2(\boldsymbol{X})]$ 称为参数 θ 的置信度（置信水平）为 $1-\alpha$ 的置信区间。 ♣

注 你很难说明一个“点估计”究竟有多么精确，但是可以通过点估计来构造置信区间，它给了你更多理由。

注 如果你只有充分把握得到置信区间的一个界，那么你得到的叫做“置信上限”或者“置信下限”。也十分有用。

这里，如果你有精力的话，可以去熟悉一些数理统计书上关于置信区间的例子，重点是正态等。至少，回归分析不会要求你掌握构造各种各样的置信区间的办法，你只需要看得懂，然后尝试着去理解。

2.2.4 参数假设检验

假设检验是数理统计学最精彩的内容，在这里你也会初次感受到统计并不是数学，更不是数数、按计算器的学科。统计是一门艺术，有人的因素在里面。

初学假设检验的时候我是十分震撼的，因此在写这一部分的时候诚惶诚恐，担心没办法把这部分当初带给我的感受充分地表达出来。

这一节我主要讲参数假设检验，非参数假设检验的一些具体方法我将在下一节里面作简要介绍。

设有一参数分布族 $\mathcal{F}=\{f(x,\theta):\theta\in\Theta\}$, 其中 Θ 为参数空间。令 $X=(X_1,\ldots,X_n)$ 为从 $\mathcal{F}$ 中抽取的简单随机样本。

这句话我写过很多遍了，可以看出研究每个问题的时候，你拥有的都是一些样本，以及对于问题的了解（对参数空间的假设、先验知识等）。

在参数假设检验问题中，感兴趣的是 θ 是否属于参数空间的某个非空真子集 Θ_0. 则命题 $H_0:\theta\in\Theta_0$ 称为零假设或原假设。命题 $\theta\in\Theta_1$ 称为对立假设或备择假设。

注 在统计中，我们一定要选取**公认的、不容易被否定的命题作为原假设**，实验者新提出的、想要说明的结论作为备择假设，这是规定。实验者的目的是推翻原假设，但是他必须有极大的把握（例如 95% 以上）才能做出“否定原假设”的结论，否则，如果他只能说明他有 70% 的概率认为原假设不对，也不能做出“否定原假设”的结论，而只能“勉强接受原假设”。当然，具体的数值还与置信度 α 有关。

注 **原假设和对立假设的地位是绝对不对等的！**在假设检验问题中我们最终做出的判断是“接受”或者“拒绝”原假设，事实上对立假设完全可以不写（视为原假设空间的补集），不管它具体为什么形式，我们做统计推断时只关心原假设的形式。

假设检验问题可以表示为

$$H_0:\theta\in\Theta_0\longleftrightarrow H_1:\theta\in\Theta_1.$$

下面看一个例子：

例题 2.2 设 $X=(X_1,\ldots,X_n)$ 为从正态总体 $N(a,\sigma^2)$ 中抽取的随机样本，其中 σ^2 已知。考虑检验问题

$$H_0: a=a_o \leftrightarrow H_1: a\neq a_0,$$

此处 a_0 为给定的常数。即我们想通过收集到的样本，做出**接受**或者**否定**“$a=a_0$”这一命题的判断。

这种检验的直观做法是，先求 a 的一个估计量，可以知道 $\bar{X}=\sum\limits_{i=1}^{n}X_i$ 是 a 的一个优良估计。若 $|\bar{X}-a_0|$ 较大，就倾向于否定 H_0. 具体地说，要确定一个数 A, 当 $|\bar{X}-a_0|>A$ 时就否定 H_0, 当 $|\bar{X}-a_0|\leq A$ 时就接受 H_0. 称

$$D=\{\boldsymbol{X}=(X_1,\ldots,X_n):|\bar{X}-a_0|>A\}$$

为否定域，或拒绝域。当样本落入否定域的时候，我们就否定原假设。

为了说明我们做出的判断的有效性，我们需要从概率的角度出发，给自己的判断找理由。

比方说，假如我们可以证明，在原假设成立的条件下，$D=\{\boldsymbol{X}=(X_1,\ldots,X_n):|\bar{X}-a_0|>A\}$ 这一事件发生的概率小于 0.05（或者是更一般的 α），我们就做出“否定原假设”的判断。

注 这里的 α 称作置信水平，它的选取也需要考虑诸多因素。

注 **这里我们需要相信一件事：小概率事件不会发生**。以此，如果我们发现了在原假设成立的情况下，所收集到的样本代表的是一个小概率事件（或者是小概率事件的子集），那我们要倾向于认为我们的“原假设”是错误的，即做出“否定原假设”的结论。这时我们做出这个“否定”，“犯错误”的概率至多是 0.05，即我们有 95% 的把握认为我们的判断是正确的。（似乎是高中时学的语言）

统计推断以样本为依据，任何统计推断都不会保证绝对的正确性，而是以一定的概率去保证推断的可靠性，为人们提供更多的想法和选择。在假设检验问题中可能会犯两种错误：

1. 原假设 H_0 为真，被否定。
2. 原假设 H_0 为假，被接受。

我们希望犯两类错误的概率越小越好，但这一点很难做到，二者通常不可兼得。因为如果我们否定更多的样本，那么第一类错误就会增加，第二类错误减少，反之也可以做出相似的推论。因此，这里涉及到一个“控制犯第一类错误概率的原则”，也叫 Neyman-Pearson 原则。

公理 2.1 (Neyman-Pearson 原则)

控制犯第一类错误的概率，即在保证犯第一类错误不超过指定数值 $\alpha\in(0,1)$（通常很小），寻找犯第二类错误尽可能小的检验。 ♡

注 这里的要求和我们最初对原假设的选取相容，即选取那些不易被否定的公认的事实。假设检验时需要做到“保护原假设”，如果我们没有充分的理由说明它是错的，那么就要接受原假设。只有把握充分大的时候（犯错误概率小于检验水平 α），才能推翻它。

注 举一个例子直观理解为何如此选择。在实际问题中，犯第一类错误和第二类错误的代价通常不相同。例如我们要说明某个新药的有效性，原假设为“新药不如旧药，要坚持使用旧药”，备择假设为“抛弃旧药，选择新药”。做检验的通常是新药的研发者。而“旧药”能投入使用一定是经过更多时间和客观事实检验过的，我们将“旧药更有效”放入“不易被否定的原假设”。而如果新药的研发者想说服病人使用新药，就必须拥有**更充分**的把握，毕竟，认为“新药和旧药差不多”并不是说服人们做出改变的理由。这里犯第一类错误（否定旧药更有效）的代价可能是因为新药不好带来生命危险，但是犯第二类错误的代价仅仅是没有使用到“有可能更有效的新药”。从这个例子也可以看出，**“保护原假设”和“控制犯第一类错误的概率”是我们做推断必须要坚守的原则**。可以再回看一下 Neyman-Pearson 原则。

如果你想在《数理统计》课取得不错的考试成绩，那参数假设检验里面的各种具体例子的计算（重点是正态）就一定要熟练，但我认为理解想法以后，具体的细节其实没那么重要，读一遍就足够了。

求解假设检验问题的一般步骤：

1. 提出原假设 H_0 和备择假设；

2. 确定检验统计量和否定域的形式，A 待定；
3. 选取适当水平 α, 利用检验统计量的分布确定临界值 A；
4. 由手中的样本算出检验统计量的具体值，与否定域比较，做出接受或者拒绝原假设 H_0 的结论。

下面简单谈一谈假设检验与置信区间的关系。两者其实通常是一回事。

如果我们得到一个置信度为 α 的假设检验，例如否定域为 $D = \{|\bar{X} - \theta| > A\}$, 那么我们相当于知道 $\mathbb{P}(|\bar{X} - \theta| > A) < \alpha$, 相对应的也就能得到参数 θ 的置信区间 $[\bar{X} - A, \bar{X} + A]$, 它的置信水平为 $1 - \alpha$. 相应地，也可以由置信区间直接导出一个对应的假设检验。

这部分的最后，介绍一个概念，叫做 p-值 (p-value). 我们手上依旧有一些简单随机样本 $\boldsymbol{X} = (X_1, \ldots, X_n)$.

定义 2.6 (p 值)

在原假设成立的情况下，$p = \mathbb{P}(\text{样本}\boldsymbol{X} = (X_1, \ldots, X_n)\text{出现的概率})$ 称作 p 值。 ♣

这也是假设检验中相当重要的概念。在上面讲的一般的选取拒绝域的假设检验问题时，我们会写一些推导，然后用查表的办法确定最终的拒绝域，但这已经是很古老的办法了。现在的 p 值通常由计算机计算得到，相当方便，同时因为对于具体的分布我们可能也没办法手算这样的概率。如果我们算出来的 p 值小于预先给定的水平 α, 那么就可以作出“否定原假设”的判断。

2.2.5 一些非参数假设检验方法

回归分析中会接触到的似乎只有 χ^2 检验，用来检验列联表的独立性和齐一性。用到的便是 Pearson 卡方统计量。

例如，统计量

$$X^2 = \frac{n(ad - bc)^2}{n_1 n_0 m_1 m_0}$$

在原假设 H_0(没有差别) 的条件下近似服从 χ_1^2 分布。

具体的细节可以参考韦来生《数理统计》6.5 节。这部分我也不是很熟悉，也只是在一些案例分析使用，并不在回归的主线里，我就不详细写了。了解这些构造出来的卡方统计量以及何时否定就可以了。

在编写这章内容时，我也重新浏览了杨老师去年的《回归分析》讲义，认为以上已经包含了所有用得到的数理统计知识。当然，具体的一些细节只看这几页总结可能还是不够的，需要同学们参考不同的资料细细品味。

2.3 R 语言编程

如果想系统性地学学 R 语言，可以参考北大李东风老师的讲义：网页链接

系统性学习可以自己看完 1-19（看不明白的直接跳过），挑感兴趣的代码复制到自己的 RStudio 上跑一下。其他的用到了再查就好，条件允许的话我认为当下热门 chatGPT 更好用，你总可以用它找到你想要的的东西。

但这门课不会要求你太多。我认为回归的实验最主要的意义在于让同学自己动手去体验一下跑数据的过程，并借此对线性模型理论和相关方法产生更深刻的理解，并不是纯粹地锻炼你的敲代码能力，毕竟这门课不叫《统计软件》。所以杨老师的实验讲义绝大部分是对实验知识的讲解，再加上几个对例题稍作改变而产生的练习题。认真完成的话会收获很多，但压力不大。

2.4 线性代数

我只提几个同学们在线性代数学习过程中可能没有特别注意的话题。具体细节课上应该都会讲到，并且在老师的课件上可以看到完整的证明与推导过程，所以理论上来讲你可以之前没学过线性代数。

事实上，我所提到的这些基本只是作为工具出现在这里，**向量空间与线性变换才是线性代数最核心的内容**。

2.4.1 方阵的谱分解

这里说的其实是线性代数中实对阵矩阵的正交相似对角化，矩阵的特征根也被称作“谱”。

定理 2.2 (谱分解)

任一 $n \times n$ 对称阵 A 可以表示为

$$A = O\Lambda O^T,$$

其中 $\Lambda = diag(\lambda_1, \ldots, \lambda_n), O = (v_1, \ldots, v_n)$ 是正交矩阵，$OO^T = O^TO = I_n$, 其中 λ_i 是 A 的特征根，v_i 是对应的模长为 1 的特征向量。

这个定理很简单。如果说线性代数是机器学习的理论基础，那这个“谱分解”定理应该是其中最闪耀的部分，其余的都是陪衬。

2.4.2 矩阵的奇异值分解

定理 2.3 (奇异值分解)

任一秩为 r 的 $n \times m$ 矩阵 A 可表示为

$$A_{n\times m} = U_{n\times r} D_{r\times r} V^T_{r\times m},$$

其中 $U^TU = V^TV = I_r, D = diag(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r})$ 的对角元称为奇异值，其中 $\lambda_1 \geq \ldots \geq \lambda_r > 0$ 为 A^TA 或 AA^T 的 r 个正特征根，$U = (u_1, \ldots, u_r)$ 的第 i 列 u_i 为 AA^T 的对应于特征根 λ_i 的特征向量，$V = (v_1, \ldots, v_r)$ 的第 i 列 v_i 为 A^TA 的对应于特征根 λ_i 的特征向量。

奇异值分解是谱分解的推广，它针对的是“任一矩阵”，在实对称半正定矩阵的情况下就是谱分解。我读过的最全面的关于奇异值分解的理论，是在李航老师的《统计学习方法》，里面介绍了各种形式的奇异值分解及其证明。另外，我读过的最丰富的关于奇异值分解的应用，其实是在杨老师的《多元分析》的课件上，后半学期的统计模型，本质上都是在做奇异值分解。《回归分析》课应该应用不多。

2.4.3 广义逆

定义“广义逆”的主要动机是在求解某些优化问题时，我们因为矩阵不可逆而得不到问题的显式解。但在具体的一些问题里，我们可以证明，在需要求逆的地方我们使用任意一个“广义逆”，都能达到想要的效果。其在线性模型里最主要的应用便是写出最小二乘的显式解。

定义 2.7 (广义逆)

任何 $A_{n\times m}$, 若 $X_{m\times n}$ 满足 $AXA = A$, X 称为 A 的广义逆，记为 A^-.

广义逆的存在性总是有的，并且我们可以说明广义逆可以表达出线性方程的解。

定理 2.4

若方程 $Ax = b$ 有解，$b \neq 0$, 则任何一个解都具有形式：$x = A^-b$.

广义逆不一定唯一，但如果我们再加些限制，便能得到存在唯一的 Moore-Penrose 广义逆。由 M-P 广义逆表示的线性方程的解，是线性方程长度最小的解。

2.4.4 欧式空间中的正交投影

我们将对称幂等阵称为投影矩阵，它的特征根为 1 或 0. 如果我们想将一个向量投影到矩阵 X 的列空间上，那相应的投影阵为 $P_X = X(X^TX)^-X^T$. 由此便能得到最小二乘的解法。

第 3 章 似然方法及其理论

本章来自 文档链接。一般来讲了解前两部分内容（KL 距离、极大似然估计）加上最后的复合原假设对应的检验就够了。中间的渐进理论基本严格，但也不够严格。

似然方法是统计学中应用最广泛的方法，似然准则是统计学的最高准则之一。总体所有未知的信息都包含在似然函数中，参数的统计推断应该基于似然函数, 这就是所谓的似然准则 (likelihood principle)。

R.A.Fisher 于 1920's 提出了极大似然方法，参数的极大似然估计使得似然函数达到极大，具有相合、渐近正态、渐近有效（最优）和不变性。除了估计参数，似然也可以用来进行假设检验、置信区间等其它统计推断，这些似然推断方法同样具有最优性。我们主要介绍经典的似然理论，即极大似然估计的性质，基于似然的检验方法包括似然比检验（LRT）、Rao's score 检验，Wald 检验的零分布。

注 似然方法的重要性不仅仅局限于经典的似然理论及其应用，多似然方法有众多拓展和应用，包括条件似然 (conditional likelihood，消除冗余参数)、部分似然 (partial likelihood, Cox 半参数模型)、剖面似然（profile likelihood）、拟似然方法 (quasi-likelihood, 只假设矩函数形式而不假设具体分布)、估计方程和广义估计方程方法 (EE：EstimatingEquation，GEE)、经验似然或非参数似然等等。

3.0 Kullack-Leibler 距离

Kullack-Leibler 距离度量了两个概率密度的差异，在似然理论中有重要应用，特别地，模型选择的 AIC 准则的构建基于该度量。

引理 3.1

假设 f, g 是两个概率密度，Kullback-Leibler （KL）距离或散度 (divergence) 定义为

$$D(g|f) = E_f\left(\log\left(\frac{f}{g}\right)\right) = \int f(x)\log\left(\frac{g(x)}{f(x)}\right),$$

其中 E_f 表示对 f 求期望。则

$$D(g|f) \geq 0 \text{ 或 } E_f \log f \geq E_f \log g.$$

♡

证明：因为 $-\log(x)$ 是凸函数，所以

$$E_f\left(-\log\left(\frac{g}{f}\right)\right) \geq -\log(E_f(\frac{g}{f})) = -\log(\int \frac{g}{f} f) = -\log(\int g) = -\log(1) = 0.$$

■

3.1 极大似然估计

本节给出一些基本定义，并从直观上讨论为什么 MLE 具有优良性质。为了简便，我们仅考虑样本是 iid 并存在概率密度的情形。独立但不同分布的情形类似。假设

$$X_1, ..., X_n \; iid \sim f(x|\theta),\; \theta \in \Theta \subset R^k,\; f \text{ 是关于某个测度的密度函数.}$$

$X_1, ..., X_n$ 的联合分布 $\prod_{i=1}^n f(x_i|\theta)$ 作为未知参数 θ 的函数称为似然函数：

$$L(\theta) = \prod_{i=1}^n f(x_i|\theta).$$

对数似然函数

$$l(\theta) = \log(L(\theta)) = \sum_{i=1}^{n} \log[f(x_i|\theta)]$$

将似然函数中的乘积变成加法，易于计算和优化，更为重要的是对数似然函数是随机变量的独立和，可直接应用大数律或中心极限定理, 比如由大数律

$$l(\theta)/n \to E(\log f(X|\theta)) = \int \log f(x|\theta) \times f(x|\theta)dx,\ n \to \infty,$$

而右端的极限与 KL 距离有关。极大似然估计 $\hat{\theta}$ 使得似然函数或等价地对数似然达到最大

$$\hat{\theta} = arg\max l(\theta).$$

为什么极大似然估计是未知参数的良好估计？假设 θ 是真参数, 任取 $\theta^* \neq \theta$（即样本来自于总体 $f(x|\theta)$ 而不是 $f(x|\theta^*)$), 则由大数定律和引理 1

$$\frac{l(\theta) - l(\theta^*)}{n} = \frac{\sum_{i=1}^{n} \log(f(x_i|\theta)/f(x_i|\theta^*)}{n} \to_{a.s.} E_\theta \log\left(\frac{f(x|\theta)}{f(x|\theta^*)}\right) = \int f(x|\theta) \log\left(\frac{f(x|\theta)}{f(x|\theta^*)}\right) dx > 0,$$

其中 E_θ 表示按照分布 $f(x|\theta)$ 求期望。上述事实说明当 $n \to \infty$ 时，几乎必然有 $l(\theta) > l(\theta^*)$，即真参数使得似然函数最大，而按照定义似然估计使得似然函数最大，所以似然估计应该等于或接近真参数。即使样本量有限，使得似然函数达到极大的点即 MLE 应该在真参数 θ 附近，即 $\hat{\theta} \to \theta$（依概率或 a.s.），这就是所谓的相合性。

对数似然函数关于参数的一阶导数、二阶导数 (Hessian 矩阵) 是似然理论中最重要的量。一阶导数称为计分 (score) 函数，负的二阶导数称为（观测）Fisher 信息阵。

定义 3.1

(计分函数和信息阵)

- 对数似然函数的一阶导数称为计分函数 (score):

$$U(\theta) = \dot{l} = \frac{\partial l(\theta)}{\partial \theta} = \frac{f'(\mathbf{x}|\theta)}{f(\mathbf{x}|\theta)},$$

- 观测信息阵 (observed Fisher information):

$$I(\theta) = -\ddot{l} = -\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta^\top},$$

- Fisher 信息阵：观测信息阵的期望

$$i(\theta) = E(I(\theta)) = E\left[-\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta^\top}\right].$$

若 $X_1, X_2, ..., X_n$ iid, 则 $i(\theta) = nE\left[-\frac{\partial^2 \log(f(x|\theta)}{\partial\theta\partial\theta^\top}\right] \triangleq ni_1(\theta)$, $i_1(\theta)$ 代表单个样本的信息。

♣

定理 3.1

正则条件下，

$$EU(\theta) = 0, var(U(\theta)) = EUU^\top = i(\theta),$$

即

$$E\left(\frac{\partial l(\theta)}{\partial\theta}\right) = 0, \quad E\left(\frac{\partial l(\theta)}{\partial\theta}\right)\left(\frac{\partial l(\theta)}{\partial\theta}\right)^\top = -E\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta^\top},$$

其中的 E 是对 $X_1, ..., X_n\ iid\ \sim f(x|\theta)$ 求期望。

♡

为了求解 MLE 使得对数似然函数 $l(\theta)$ 达到极大，通常求解如下似然方程

$$U(\theta) = \partial l/\partial\theta = 0.$$

该方程未必有唯一解。

Remark1：

$i(\theta) = var(U)$ 说明 Fisher 信息阵是正定的，由大数定律，

$$\frac{1}{n}I(\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log(f(x_i|\theta)}{\partial\theta\partial\theta^\top} \to E\left(\frac{\partial^2 \log(f(x_1|\theta)}{\partial\theta\partial\theta^\top}\right) = i_1(\theta) = i(\theta)/n > 0,$$

因此当 $n \to \infty$，几乎必然 $I(\theta) > 0$. 即对数似然函数的二阶导数负定，因此似然方程几乎必然有唯一解。

Remark2：

第一个性质 $E_\theta(U(\theta)) = 0$ 说明 $U(\theta)$ 在平均意义下在真参数 θ 处为 0，或者可认为 $U(\theta) \approx 0$, 这保证了似然方程 $U(\hat{\theta}) = 0$ 的解 $\hat{\theta} \approx \theta$。

Remark3：

由 Taylor 展开

$$0 = U(\hat{\theta}) \approx U(\theta) - I(\theta)(\hat{\theta} - \theta) \approx U(\theta) - i(\theta)(\hat{\theta} - \theta)$$

知

$$\hat{\theta} - \theta \approx i(\theta)^{-1}U(\theta),$$

因为 $U = U(\theta)$ 是独立随机变量之和，定理 1 求出了 U 的均值和方差，则由中心极限定理

$$var(U)^{-1/2}\left(U - E(U)\right) = i(\theta)^{-1/2}U(\theta) = i_1^{-1/2}(\theta)U(\theta)/\sqrt{n} \xrightarrow{d} N(0, I_k).$$

即

$$U(\theta)/\sqrt{n} \xrightarrow{d} N(0, i_1(\theta)).$$

由上述两式可得 $\hat{\theta}$ 服从渐近正态分布

$$\sqrt{n}(\hat{\theta} - \theta) = \sqrt{n}i(\theta)^{-1}U(\theta) = i_1^{-1}U(\theta)/\sqrt{n} \xrightarrow{d} N(0, i_1(\theta)^{-1}).$$

以上各个 Remark 从直观上大概说明了 MLE 应该具有的优良性。下面将我们上述直观（稍微）严格化。

3.2 极大似然估计的渐近性质

正则条件 (Regularity Conditions)

i. f 的支撑不依赖于 θ.

ii. f 是可识别的（若 $f(x|\theta_1) = f(x|\theta_2)$ 对任何 x 都成立，则必有 $\theta_1 = \theta_2$），$\Theta \subset R^k$ 是开集.

iii. $\left|\frac{\partial^3 \log(f(x|\theta))}{\partial\theta^3}\right| \le M(x)$，$E_\theta M(x) < \infty$.

将 $U(\theta)$ 在真参数 θ 处 Taylor 展开并利用 $I(\theta)/n \to i(\theta)$,

$$0 = U(\hat{\theta}) = U(\theta) + -I(\theta)(\hat{\theta} - \theta) + o_p(n^{-\frac{1}{2}}) = U(\theta) - ni_1(\theta)(\hat{\theta} - \theta) + o_p(n^{-\frac{1}{2}})$$

所以我们可将 MLE 表示为（为了简单，省略小 o 项）

$$\hat{\theta} - \theta = [ni_1(\theta)]^{-1}U(\theta). \tag{3.1}$$

or

$$\sqrt{n}(\hat{\theta} - \theta) = i_1^{-1}(\theta)\frac{U(\theta)}{\sqrt{n}}. \tag{3.2}$$

下面将多次使用该表示。

定理 3.2

(相合性 consistency) 正则条件下，MLE $\hat{\theta}$ 是参数 θ 的相合估计，即

$$\hat{\theta} \xrightarrow{Pr.} \theta.$$

Proof. 由大数定律，$U(\theta)/n = \sum_{i=1}^{n} \frac{\partial \log f(x_i|\theta)}{\partial \theta} \to E(\frac{\partial \log f(x_1|\theta)}{\partial \theta}) = 0$, 所以 $\hat{\theta} = \theta + i^{-1}(\theta)\frac{U}{n} + o_p(1) \to 0\ a.s.$

定理 3.3

(渐近正态 Asymptotic normality) 正则条件下，

$$\frac{U(\theta)}{\sqrt{n}} \to_d N(0, i_1(\theta)), \tag{3.3}$$

$$\sqrt{n}(\hat{\theta} - \theta) \to_d N(0, i_1(\theta)^{-1}). \tag{3.4}$$

Proof. 注意到 score 函数 $U(\theta)$ 是独立随机变量的和，$E(U(\theta)) = 0, var(U(\theta)) = i(\theta) = ni_1(\theta)$, 由中心极限定理

$$var(U)^{-1/2}[U(\theta) - EU(\theta)] = (ni_1(\theta))^{-1/2}U(\theta) \xrightarrow{d} N(0, I_k),$$

即

$$U(\theta)/\sqrt{n} \xrightarrow{d} N(0, i_1(\theta)).$$

由（2）

$$\sqrt{n}(\hat{\theta} - \theta) = i_1^{-1}(\theta)\frac{U(\theta)}{\sqrt{n}} \xrightarrow{d} N(0, i_1(\theta)^{-1}).$$

■

极大似然估计是渐近最优的，即在极限情况下，在所有 θ 的无偏或渐近无偏估计中，极大似然估计的方差最小。

定理 3.4

(渐近最优 Asymptotic efficiency,Hajek) 若 $\tilde{\theta}$ 是任一 θ 的无偏估计, 则它可以分解为

$$\tilde{\theta} = \hat{\theta} + Z$$

其中 Z 与 $\hat{\theta}$ 不相关，这表明

$$var(\tilde{\theta}) = var(\hat{\theta}) + var(Z) \geq var(\hat{\theta}).$$

Proof. $dim(\theta) = k$，对任何无偏检验 $\tilde{\theta}, \theta = E_\theta\tilde{\theta} = \int \tilde{\theta}L(\theta)$, 两边同时对 θ 求导，得 (注意似然函数 L 是联合概率密度)

$$I_k = \int \tilde{\theta}L' = \int \tilde{\theta}L'/L \times L = E\tilde{\theta}L'/L = E(\tilde{\theta}U(\theta))^\top,$$

其中 $L'/L = (\log L)' = U(\theta)$, 所以

$$E(\tilde{\theta} - \hat{\theta})U^\top = I_k - E(\hat{\theta}U^\top)U = I_k - E(i^{-1}UU^\top) = I_k - I_k = 0,$$

所以 $Z = \tilde{\theta} - \hat{\theta}$ 与 $\perp \hat{\theta}$ 不相关，

$$\Rightarrow \tilde{\theta} = \hat{\theta} + Z, \text{ where } Z \perp \hat{\theta}.$$

下述定理可用于构造假设检验和置信区间

定理 3.5

正则条件下，下述三个量渐近等价且都收敛到卡方分布

$$2(l(\hat{\theta}) - l(\theta)) \approx (\hat{\theta} - \theta)^\top[ni_1(\theta)](\hat{\theta} - \theta) \approx U(\theta)^\top[ni_1(\theta)]^{-1}U(\theta) \to_d \chi_k^2, k = dim(\theta) \tag{3.5}$$

因为 $I(\theta)/n \to i_1(\theta)$, 由 Slusky 引理，上式中 $i(\theta) = ni_1(\theta)$ 换成 $I(\theta)$ 仍然成立。 ♡

Proof. 二阶 Taylor 展开,

$$l(\hat{\theta}) - l(\theta) = U(\theta)^\top I(\hat{\theta}-\theta) - \frac{1}{2}(\hat{\theta}-\theta)^\top I(\theta)(\hat{\theta}-\theta) + o_p(1) = U(\theta)^\top(\hat{\theta}-\theta) - \frac{1}{2}\sqrt{n}(\hat{\theta}-\theta)^\top i_1(\theta)\sqrt{n}(\hat{\theta}-\theta) + o_p(1)$$

由 (3.1), $U(\theta) = ni_1(\theta)(\hat{\theta}-\theta)$, 上述 Taylor 展开中以 $ni_1(\theta)(\hat{\theta}-\theta)$ 替代 $U(\theta)$，得

$$l(\hat{\theta}) - l(\theta) = \frac{1}{2}(\hat{\theta}-\theta)^\top[ni(\theta)](\hat{\theta}-\theta) + o_p(1)$$

Talyor 展开中以 $[ni_1(\theta)]^{-1}U(\theta)$ 替代 $\hat{\theta}-\theta$，得

$$l(\hat{\theta}) - l(\theta) = \frac{1}{2}U(\theta)^\top[n_1 i(\theta)]^{-1}U(\theta)$$

由定理 3 知近似地有 $U(\theta) \sim N(0, ni_1(\theta))$，$\sqrt{n}(\hat{\theta}-\theta) \sim N(0, (ni_1(\theta))^{-1})$，从而上述各个量都依分布收敛到 χ_k^2。 ■

3.3 基于似然的检验：简单原假设情形

考虑简单原假设 $H_0 : \theta = \theta_0$。基于定理 5 的结论，可构造基于似然的三种常见检验。

- 似然比检验 (LRT)

$$LRT = -2\log(\lambda) = 2[l(\hat{\theta}) - l(\theta_0)] \sim_{H_0} \chi_k^2,$$

其中似然比 $\lambda = L(\theta_0)/L(\hat{\theta})$。

- Wald 检验

$$W = (\hat{\theta}-\theta_0)^\top I(\theta_0)(\hat{\theta}-\theta_0) \sim_{H_0} \chi_k^2.$$

- Score 检验

$$S = U(\theta_0)^\top I(\theta_0)^{-1}U(\theta_0) \sim_{H_0} \chi_k^2.$$

当上述检验统计量大于 $\chi_k^2(\alpha)$ 时，在水平 α 下拒绝 H_0。三个统计量渐近等价。W, S 中的 $I(\theta_0)$ 可换成 $ni_1(\theta_0)$(后者需要更多的计算，需要计算期望，因而不太常用，事实上其功效略差于前者)。

Remark4:

三个检验渐近等价，一般具有相近的功效，但它们在具体应用中还是有差异的。LRT 计算最简单，无需计算对数似然的导数；Wald 检验最直观，它直接对比 $\hat{\theta}$ 和 θ_0 的差异；Score 检验考察 $U(\theta_0)$(的模) 是否接近 0，若接近 0，那么 $\hat{\theta} \approx \theta_0$ (因为 $U(\hat{\theta}) = 0$).，从而不能拒绝原假设。

3.4 基于似然的检验：复合原假设情形

复合假设是更常见的情形。假设参数空间 $\Theta \subset R^k$,$dim(\Theta) = k$。我们希望检验复合原假设

$$H_0 : \theta \in \Theta_0 \subset \Theta, \quad dim(\Theta_0) = m < k.$$

假设存在二次可导的函数 g 使得 $\Theta_0 = g(R^m)$，则在正则条件下，当原假设成立时

$$-2\left[\log\left(\frac{L(\hat{\theta}_0)}{L(\hat{\theta})}\right)\right] \to_d \chi_{k-m}^2 \text{ as } n \to \infty,$$

其中

$$\hat{\theta}_0 = \arg\max_{\theta\in\Theta_0} l(\theta); \hat{\theta} = \arg\max_{\theta\in\Theta} l(\theta).$$

通过重新参数化，上述复合假设问题可转化为如下情形（实践中最常见的情形），即参数 θ 可划分为两部分：

$$\theta_{k\times 1} = (\psi_{(k-m)\times 1}, \lambda_{m\times 1})^\top$$

其中我们感兴趣的或者需要检验的是参数 ψ，而 λ 称为冗余参数（nuisance parameter)。原假设为

$$H_0 : \psi = \psi_0 \ \ (\psi_0 \text{ 已知})$$

原假设成立时的参数空间为

$$\Theta_0 = \{(\psi_0, \lambda) : \lambda \in R^m\}, \ \ dim(\Theta_0) = m.$$

对数似然函数

$$l(\theta) = l(\psi, \lambda) = \sum_{i=1}^{n} \log f(x_i; \theta)$$

记 MLE $\hat{\theta} = (\hat{\psi}, \hat{\lambda})^\top$, 原假设成立时 ψ 的 MLE 为 $\hat{\psi}_0$, 记 $\hat{\theta}_0 = (\psi_0, \hat{\lambda}_0)^\top$. 划分计分函数

$$U(\theta) = \frac{\partial l}{\partial \theta} = \begin{pmatrix} \frac{\partial l}{\partial \psi} \\ \frac{\partial l}{\partial \lambda} \end{pmatrix} \triangleq \begin{pmatrix} U_\psi \\ U_\lambda \end{pmatrix}$$

划分观测信息阵如下

$$I = -\frac{\partial^2 l}{\partial \theta \partial \theta^\top} = \begin{pmatrix} I_{\psi\psi} & I_{\psi\lambda} \\ I_{\lambda\psi} & I_{\lambda\lambda} \end{pmatrix}$$

记其逆矩阵

$$I^{-1} = \begin{pmatrix} I^{\psi\psi} & I^{\psi\lambda} \\ I^{\lambda\psi} & I^{\lambda\lambda} \end{pmatrix},$$

其中 $I^{\psi\psi} = I_{\psi\psi\bullet\lambda} = I_{\psi\psi} - I_{\psi\lambda} I_{\lambda\lambda}^{-1} I_{\lambda\psi}$。类似于简单假设情形，我们定义似然比检验（LRT），score 检验 S, Wald 检验 W 如下。

复合假设的似然检验

- 似然比检验

$$LRT = 2[l(\hat{\theta}) - l(\hat{\theta}_0)]$$

- Score 检验

$$S = U_\psi(\hat{\theta}_0)^\top I_{\psi\psi}^{-1}(\hat{\theta}_0) I_\psi(\hat{\theta}_0) \ \ (\hat{\theta}_0 : \text{原假设下的 MLE})$$

- Wald 检验

$$W = (\hat{\psi} - \psi_0)^\top I^{\psi\psi}(\hat{\theta})^{-1}(\hat{\psi} - \psi_0) \ \ (\hat{\theta} : \text{无假设下的 MLE})$$

三个检验在渐近等价，原假设成立时它们的极限分布都是 χ^2_{k-m}(定理 6) 。

Remark5：

Remark4 仍适用于这里。另外，许多著名的检验统计量都是复合情形下的 Score 检验（比如列联表的 Pearson 卡方法检验）。

例 1. 假设 I 个多项分布总体

$$\mathbf{x}_i = (x_{i1}, ..., x_{iJ})^\top \sim Multinomial(n_i; \mathbf{p}_i),$$

其中 $\mathbf{p}_i = (p_{i1}, ..., p_{iJ})^\top, p_{i1} + ... + p_{iJ} = 1, \ p_{ij} > 0, i = 1, ..., J$, 将所有计数组成 $I \times J$ 列联表 $X = (x_{ij}, 1 \le i \le I, 1 \le j \le J)$。齐一性假设

$$H_0 : \mathbf{p}_1 = ... = \mathbf{p}_I.$$

则 H_0 的 score 检验为列联表的 Pearson 卡方检验。

定理 3.6

LRT, S, W 渐近等价，原假设 $H_0: \psi = \psi_0$ 成立时它们的分布收敛到 χ^2_{k-m}，$n \to \infty$。 ♡

Appendix: Sketch proof of Theorem 6

We only prove LRT. Let $\hat{\theta}_0 = (\psi_0, \hat{\lambda}_0)$, where $\hat{\lambda}_0$ is the MLE under the null, i.e, we have

$$\frac{\partial l(\theta)}{\partial \lambda} |_{\psi=\psi_0, \lambda=\hat{\lambda}_0} = 0$$

The global MLE is denoted by $\hat{\theta} = (\hat{\psi}, \hat{\lambda})$, then by Taylor expansion, the log-likelihood ratio is (omitting high-order terms)

$$\begin{aligned} l(\hat{\theta}_0) - l(\hat{\theta}) &= \frac{\partial l(\theta)}{\partial \theta}|_{\theta=\hat{\theta}} + \frac{1}{2}(\hat{\theta}_0 - \hat{\theta})^\top \left[\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta^\top} |_{\theta=\hat{\theta}}\right](\hat{\theta}_0 - \hat{\theta}) \\ &\approx -\frac{n}{2}(\hat{\theta}_0 - \hat{\theta})^\top \left[i(\hat{\theta})\right](\hat{\theta}_0 - \hat{\theta}) \end{aligned} \tag{3.6}$$

where we have subsituted $\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta^\top} |_{\theta=\hat{\theta}}$ by its expectation at true parameter $\theta_0 = (\psi_0, \lambda)$.

Next, we expand $U(\hat{\theta}_0)$ around $\hat{\theta}$.

$$\begin{aligned} U(\hat{\theta}_0) &= \begin{pmatrix} \frac{\partial l(\theta)}{\partial \psi} |_{(\psi_0, \hat{\lambda}_0)} \\ \frac{\partial l(\theta)}{\partial \lambda} |_{(\psi_0, \hat{\lambda}_0)} \end{pmatrix} = \begin{pmatrix} \frac{\partial l(\theta)}{\partial \psi} |_{(\psi_0, \hat{\lambda}_0)} \\ 0 \end{pmatrix} \\ &= \frac{\partial U}{\partial \theta}(\hat{\theta}_0 - \hat{\theta}) = -ni(\theta_a)(\hat{\theta}_0 - \hat{\theta}) = -n \begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\lambda\psi} & i_{\lambda\lambda} \end{pmatrix} \begin{pmatrix} \psi_0 - \hat{\psi} \\ \hat{\lambda}_0 - \hat{\lambda} \end{pmatrix} \\ &= -n \begin{pmatrix} i_{\psi\psi}(\psi_0 - \hat{\psi}) + i_{\psi\lambda}(\hat{\lambda}_0 - \hat{\lambda}) \\ i_{\lambda\psi}(\psi_0 - \hat{\psi}) + i_{\lambda\lambda}(\hat{\lambda}_0 - \hat{\lambda}) \end{pmatrix} \end{aligned}$$

So from the second component in the above equation we have

$$\hat{\lambda}_0 - \hat{\lambda} = -i^{-1}_{\lambda\lambda} i_{\lambda\psi}(\psi_0 - \hat{\psi})$$

Then,

$$\hat{\theta}_0 - \hat{\theta} = \begin{pmatrix} \psi_0 - \hat{\psi} \\ \hat{\lambda}_0 - \hat{\lambda} \end{pmatrix} = \begin{pmatrix} \psi_0 - \hat{\psi} \\ -i^{-1}_{\lambda\lambda} i_{\lambda\psi}(\psi_0 - \hat{\psi}) \end{pmatrix} = \begin{pmatrix} I \\ -i^{-1}_{\lambda\lambda} i_{\lambda\psi} \end{pmatrix} \left(\psi_0 - \hat{\psi}\right) \tag{3.7}$$

Then from (6) and (7), we have

$$\begin{aligned} l(\hat{\theta}_0) - l(\hat{\theta}) &\approx -\frac{n}{2}(\psi_0 - \hat{\psi})^\top (I, -i_{\psi\lambda} i^{-1}_{\lambda\lambda}) \begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\lambda\psi} & i_{\lambda\lambda} \end{pmatrix} \begin{pmatrix} I \\ -i^{-1}_{\lambda\lambda} i_{\lambda\psi} \end{pmatrix} \left(\psi_0 - \hat{\psi}\right) \\ &= -\frac{n}{2}(\psi_0 - \hat{\psi})^\top i_{\psi\psi\bullet\lambda}(\theta_0)(\psi_0 - \hat{\psi}), \end{aligned} \tag{3.8}$$

where $i_{\psi\psi\bullet\lambda} = i_{\psi\psi} - i_{\psi\lambda} i^{-1}_{\lambda\lambda} i_{\lambda\lambda}$. Since under the null hypothesis, true parameter is $\theta_0 = (\psi_0, \lambda)^\top$, $\sqrt{n}(\hat{\theta} - \theta_0) \to_d N(0, i^{-1}(\theta_0))$, then marginally

$$\sqrt{n}(\hat{\psi} - \psi_0) \to N(0, i^{-1}_{\psi\psi\bullet\lambda}(\theta_0)),$$

combining with (8), we have

$$-2\left\{l(\hat{\theta}_0) - l(\hat{\theta})\right\} = \left[\sqrt{n}(\hat{\psi} - \psi_0)\right] [i_{\psi\psi\cdot\lambda}(\theta_0)] \left[\sqrt{n}(\hat{\psi} - \psi_0)\right] \to \chi^2_{k-m}$$

where $k - m$ is the length of ψ.

Reference：

Cox and Hinkley (1974) Theoretical Statistics. CRC.

第 4 章 作业解答

4.1 HW 1

作业 1 链接
作业 1 解答

4.2 HW 2

作业 2 链接

练习 4.1 随机向量 $\mathbf{x}=(x_1,\ldots,x_n)^\top$, 假设随机变量 $x_1,\ldots,x_n$ iid $\sim N(0,1)$, 我们称随机向量 $\mathbf{x}$ 服从 n 元标准正态分布, 记作 $\mathbf{x}\sim N_n(0,I_n)$ 。假设 A 是 $n\times n$ 正交矩阵, 证明 $\mathbf{y}=A\mathbf{x}\sim N_n(0,I_n)$.

证明 设 $A=(a_{ij})_{n\times n}$

$\because A$ 是正交阵 $\quad\therefore \sum_{j=1}^n a_{ij}^2=1\quad \forall i,\quad \sum_{j=1}^n a_{kj}a_{ij}=0\quad \forall k\neq l$

$\because Y=\begin{pmatrix} y_1\\ \vdots\\ y_n\end{pmatrix}=Ax\quad y_i=\sum_{j=1}^n a_{ij}x_j$. 而 $x_1,\cdots x_n\overset{\text{iid}}{\sim}N(0,1)$

$\therefore y_i\sim N(0,1)$

又对 $\forall k\neq l\quad \operatorname{Cov}(y_k,y_l)=\sum_{j=1}^n a_{kj}a_{lj}=0$, 即 y_k 与 y_l 独立

$$\therefore Y=Ax\sim N_n(0,I_n)$$

练习 4.2 设 $\mathbf{a},\mathbf{b}\in R^n$ 为常数向量, 假设 $\|\mathbf{a}\|=\|\mathbf{b}\|=1$, 且 $\mathbf{a}^\top\mathbf{b}=0$ 。假设 $n\times 1$ 随机向量 $\mathbf{x}\sim N_n(0,I_n)$, 试证明

$$\sqrt{n-2}\times\frac{\mathbf{a}^\top\mathbf{x}}{\sqrt{\|\mathbf{x}\|^2-(\mathbf{a}^\top\mathbf{x})^2-(\mathbf{b}^\top\mathbf{x})^2}}\sim t_{n-2}.$$

证明 构造 $n\times n$ 的正交矩阵 $A=\begin{pmatrix} a^\top\\ b^\top\\ *\end{pmatrix}$, 令 $y=Ax=\begin{pmatrix} a^\top x\\ b^\top x\\ *\end{pmatrix}=\begin{pmatrix} y_1\\ \vdots\\ y_n\end{pmatrix}$ 由题 1 可知, $y_1,\cdots y_n\overset{\text{iid}}{\sim}N(0,1)$

即 $y\sim N_n(0,I_n)$

$$\because \|y\|^2=y^\top y=x^\top A^\top Ax=x^\top I_n x=x^\top x=\|x\|^2$$

$$\therefore \|x\|^2-\left(a^\top x\right)^2-\left(b^\top x\right)^2=\|y\|^2-y_1^2-y_2^2=\sum_{i=3}^n y_i^2\sim\chi_{n-2}^2$$

$$\therefore \sqrt{n-2}\frac{a^\top x}{\sqrt{\|x\|^2-(a^\top x)^2-(b^\top x)^2}}=\frac{y_1}{\sqrt{\sum_{i=3}^n y_i^2/(n-2)}}\sim t_{n-2}$$

练习 4.3 假设两组样本:

$$y_1,y_2,\ldots,y_{n_0}\text{ iid }\sim N\left(\mu_0,\sigma^2\right),$$

$$y_{n_0+1},y_{n_0+2},\ldots,y_{n_0+n_1}\text{ iid }\sim N\left(\mu_1,\sigma^2\right).$$

零假设 $H_0:\mu_1=\mu_0$ 的两样本 t-检验统计量为

$$t_1=\frac{\bar y_1-\bar y_0}{\sqrt{\left(\frac{1}{n_0}+\frac{1}{n_1}\right)s^2}},$$

其中 $\bar y_0=\frac{1}{n_0}\sum_{i=1}^{n_0}y_i,\bar y_1=\frac{1}{n_1}\sum_{i=n_0+1}^{n_0+n_1}y_i,s^2=\frac{1}{n_0+n_1-2}\left(\sum_{i=1}^{n_0}(y_i-\bar y_0)^2+\sum_{i=n_0+1}^{n_0+n_1}(y_i-\bar y_1)^2\right)$ 。另外一方面, 我们记 $x_i=0,1\le i\le n_0$ 为第一组的标号, $x_i=1,n_0+1\le i\le n_0+n_1$ 为第二组的标号。记 r 为 $(x_i,y_i),i=1,\ldots,n=n_0+n_1$ 的样本相关系数, x,y 的相关性检验统计量

$$t_2=\sqrt{n-2}\frac{r}{\sqrt{1-r^2}},$$

试验证 $t_1=t_2$ 。

证明

$$
\begin{aligned}
&r=\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}},\quad \bar{x}=\frac{n_1}{n_0+n_1},\quad \bar{y}=\frac{n_0\bar{y}_0+n_1\bar{y}_1}{n_0+n_1}\\
&S_{xx}=\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2=n_0\left(\frac{n_1}{n_0+n_1}\right)^2+n_1\left(\frac{n_0}{n_0+n_1}\right)^2=\frac{n_1n_0}{n_1+n_0}\\
&S_{xy}=\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)\\
&=n_0\left(-\frac{n_1}{n_0+n_1}\right)\left(\bar{y}_0-\bar{y}\right)+n_1\left(\frac{n_0}{n_0+n_1}\right)\left(\bar{y}_1-\bar{y}\right)\\
&=\frac{n_0n_1^2\left(\bar{y}_1-\bar{y}_0\right)}{\left(n_0+n_1\right)^2}+\frac{n_0^2n_1\left(\bar{y}_1-\bar{y}_0\right)}{\left(n_0+n_1\right)^2}\\
&=\frac{n_0n_1\left(\bar{y}_1-\bar{y}_0\right)}{n_0+n_1}\\
&S_{yy}=\sum_{i=1}^{n}\left(y_i-\bar{y}\right)=\sum_{i=1}^{n}y_i^2-n\bar{y}^2=\left(\sum_{i=1}^{n}y_i^2-n_0\bar{y}_0^2-n_1\bar{y}_1^2\right)+\left(n_0\bar{y}_0^2+n_1\bar{y}_1^2-n\bar{y}^2\right)\\
&=\left(n_0+n_1-2\right)s^2+\frac{n_0n_1}{n_0+n_1}\left(\bar{y}_0-\bar{y}_1\right)^2\\
&r=\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}=\sqrt{\frac{n_0n_1}{n_0+n_1}}\frac{\bar{y}_1-\bar{y}_0}{\sqrt{S_{yy}}},1-r^2=1-\frac{n_0n_1}{n_0+n_1}\frac{\left(\bar{y}_1-\bar{y}_0\right)^2}{S_{yy}}=\frac{\left(n_0+n_1-2\right)s^2}{S_{yy}}\\
&\therefore\frac{r}{\sqrt{1-r^2}}=\sqrt{\frac{n_0n_1}{n_0+n_1}}\frac{\bar{y}_1-\bar{y}_0}{\sqrt{(n_0+n_1-2)s^2}}=\frac{\bar{y}_1-\bar{y}_0}{\sqrt{\left(\frac{1}{n_0}+\frac{1}{n_1}\right)s^2}}\cdot\frac{1}{\sqrt{n-2}}\\
&\therefore\sqrt{n-2}\cdot\frac{r}{\sqrt{1-r^2}}=t_1
\end{aligned}
$$

练习 4.4 假设有二元数据 $(x_i,y_i),i=1,\ldots,n$iid, 其中 x_i,y_i 各分别仅取 0,1 值, 其样本相关系数记为 r 。另一方面, 这种二元属性 (离散) 数据常常以列联表的形式表示: 其中 a 为 $x_i=y_i=1$ 的样本个数 (即 $a=\sum x_iy_i,c=\sum(1-x_i)y_i$ 等等), $n_1=a+b,n_0=c+d,m_1=a+c,m_0=b+d$ 。该列联表的独立性检验的 Pearson 卡方统计量为

$$X^2=\frac{n(ad-bc)^2}{n_0n_1m_0m_1}.$$

试验证等于 $X^2=nr^2$ 。

证明

$$
\begin{aligned}
&r^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} \\
&S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = a - \frac{n_1 m_1}{n} \\
&S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = n_1 - \frac{n_1^2}{n} \\
&S_{yy} = m_1 - \frac{m_1^2}{n} \\
&\therefore r^2 = \frac{\left(a - \frac{n_1 m_1}{n}\right)^2}{\left(n_1 - \frac{n_1^2}{n}\right)\left(m_1 - \frac{m_1^2}{n}\right)} = \frac{(na - n_1 m_1)^2}{n_0 n_1 m_1 m_0} \\
&= \frac{(ad - bc)^2}{n_0 n_1 m_0 m_1} \\
&\therefore X^2 = nr^2
\end{aligned}
$$

4.3 HW 3

4.3.1 习题

作业 3 链接

记号: 对于任何随机变量或向量 a, b, c, 我们记 $\Sigma_{ab} = \mathrm{cov}(a, b), \Sigma_{ab\bullet c} = \Sigma_{ab} - \Sigma_{ac}\Sigma_{cc}^{-1}\Sigma_{cb}$

练习 4.1 假设 $n \times 1$ 随机向量 $\mathbf{x} = (x_1, ..., x_n)^\top \sim (\boldsymbol{\mu}, \Sigma)$，即 $E(\mathbf{x}) = \boldsymbol{\mu}, \mathrm{cov}(\mathbf{x}) = \Sigma$，设 $\mathbf{a}, \mathbf{b} \in R^n$ 为任意两个常数向量

1. 求随机变量 $\mathbf{a}^\top\mathbf{x}$ 的均值和方差;
2. 求 $(x_1 + ... + x_n)/n$ 的均值和方差;
3. 求 $\mathbf{a}^\top\mathbf{x}$ 和 $\mathbf{b}^\top\mathbf{x}$ 的协方差和相关系数 $\rho(\mathbf{a}^\top\mathbf{x}, \mathbf{b}^\top\mathbf{x})$。
4. 求 $E||\mathbf{x}||^2$.

解

1. $E\boldsymbol{a}^T\boldsymbol{x} = \boldsymbol{a}^T\mu, Var(\boldsymbol{a}^T\boldsymbol{x}) = cov(\boldsymbol{a}^T\boldsymbol{x}, \boldsymbol{a}^T\boldsymbol{x}) = \boldsymbol{a}^T\Sigma\boldsymbol{a}$.
2. $E(\frac{x_1+...+x_n}{n}) = \mu, var(\frac{x_1+...+x_n}{n}) = \boldsymbol{a}^T\Sigma\boldsymbol{a}$, where $\boldsymbol{a} = \frac{1}{n}\mathbf{1}$.
3. $cov(\boldsymbol{a}^Tx, \boldsymbol{b}^Tx) = \boldsymbol{a}^T\Sigma\boldsymbol{b}, \rho(\boldsymbol{a}^Tx, \boldsymbol{b}^Tx) = \frac{cov(\boldsymbol{a}^Tx, \boldsymbol{b}^Tx)}{\sqrt{Var(\boldsymbol{a}^T\boldsymbol{x})Var(\boldsymbol{b}^T\boldsymbol{x})}} = \frac{\boldsymbol{a}^T\Sigma\boldsymbol{b}}{\sqrt{\boldsymbol{a}^T\Sigma\boldsymbol{a}\boldsymbol{b}^T\Sigma\boldsymbol{b}}}$.
4. $E\|x\|^2 = \Sigma_{i=1}^n Ex_i^2 = \Sigma_{i=1}^n(Var(x_i) + (Ex_i)^2) = tr\Sigma + \|\mu\|^2$.

练习 4.2 假设 $\mathbf{y}, \mathbf{x}, \mathbf{z}$ 为三个随机向量，记它们的方差-协方差矩阵为

$$\Sigma = \mathrm{var}\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \\ \mathbf{z} \end{pmatrix} = \begin{pmatrix} \Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{xy}} & \Sigma_{\mathbf{xz}} \\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{yy}} & \Sigma_{\mathbf{yz}} \\ \Sigma_{\mathbf{zx}} & \Sigma_{\mathbf{zy}} & \Sigma_{\mathbf{zz}} \end{pmatrix}$$

令 $\mathbf{x}, \mathbf{y}$ 关于 $\mathbf{y}$ 的去相关化分别为

$$\mathbf{x}^\perp = \mathbf{x} - \Sigma_{\mathbf{xz}}\Sigma_{\mathbf{zz}}^{-1}\mathbf{z},\ \mathbf{y}^\perp = \mathbf{y} - \Sigma_{\mathbf{yz}}\Sigma_{\mathbf{zz}}^{-1}\mathbf{z}$$

验证

$$\mathrm{var}\begin{pmatrix} \mathbf{x}^\perp \\ \mathbf{y}^\perp \\ \mathbf{z} \end{pmatrix} = \begin{pmatrix} \Sigma_{\mathbf{xx}\bullet\mathbf{z}} & \Sigma_{\mathbf{xy}\bullet\mathbf{z}} & 0 \\ \Sigma_{\mathbf{yx}\bullet\mathbf{z}} & \Sigma_{\mathbf{yy}\bullet\mathbf{z}} & 0 \\ 0 & 0 & \Sigma_{\mathbf{zz}} \end{pmatrix}$$

证明 只要用 $\Sigma_{\boldsymbol{xy}\bullet\boldsymbol{z}}$ 的定义验证

$$var(\boldsymbol{x}^\perp) = \Sigma_{\boldsymbol{xx}\bullet\boldsymbol{z}}$$
$$var(\boldsymbol{y}^\perp) = \Sigma_{\boldsymbol{yy}\bullet\boldsymbol{z}}$$
$$cov({\boldsymbol{x}^\perp}^T\boldsymbol{y}^\perp) = \Sigma_{\boldsymbol{xy}\bullet\boldsymbol{z}}.$$

由去相关化定义，

$$cov({\boldsymbol{x}^\perp}^T\boldsymbol{z}) = 0,$$

整合成矩阵形式可知原命题成立。

练习 4.3 假设 y 是随机变量, $\mathbf{x}$ 是 $p \times 1$ 随机向量, 对任何常数向量 $\mathbf{a} \in R^p$，证明 $\mathbf{a}^\top\mathbf{x}$ 和 y 的相关系数 $\rho(\mathbf{a}^\top\mathbf{x}, y) = \frac{\mathbf{a}^\top\Sigma_{\mathbf{x}y}}{\sqrt{\mathbf{a}^\top\Sigma_{\mathbf{xx}}\mathbf{a}}\sqrt{\Sigma_{yy}}}$, 且 $|\rho(\mathbf{a}^\top\mathbf{x}, y)| \le \sqrt{\frac{\Sigma_{y\mathbf{x}}\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{x}y}}{\Sigma yy}}$.

注 希望同学们通过本题，熟悉矩阵形式的 Cauchy-Schwarz 不等式，并熟练运用配凑的技术。即下面证明的第三个等号和不等式。

证明

$$
\begin{aligned}
\rho(\boldsymbol{a}^T\boldsymbol{x}, y) &= \frac{cov(\boldsymbol{a}^T\boldsymbol{x}, y)}{\sqrt{Var(\boldsymbol{a}^T\boldsymbol{x})Var(y)}} \\
&= \frac{\boldsymbol{a}^T\Sigma_{\boldsymbol{x}y}}{\sqrt{\boldsymbol{a}^T\Sigma_{\boldsymbol{x}\boldsymbol{x}}\boldsymbol{a}\Sigma_{yy}}} \\
&= \frac{(\Sigma_{\boldsymbol{x}\boldsymbol{x}}^{1/2}\boldsymbol{a})^T(\Sigma_{\boldsymbol{x}\boldsymbol{x}}^{-1/2}\Sigma_{\boldsymbol{x}y})}{\sqrt{\boldsymbol{a}^T\Sigma_{\boldsymbol{x}\boldsymbol{x}}\boldsymbol{a}\Sigma_{yy}}} \\
&\overset{C-S}{\leq} \frac{\sqrt{\boldsymbol{a}^T\Sigma_{\boldsymbol{x}\boldsymbol{x}}\boldsymbol{a}}\sqrt{\Sigma_{\boldsymbol{x}y}\Sigma_{\boldsymbol{x}\boldsymbol{x}}^{-1}\Sigma_{y\boldsymbol{x}}}}{\sqrt{\boldsymbol{a}^T\Sigma_{\boldsymbol{x}\boldsymbol{x}}\boldsymbol{a}\Sigma_{yy}}} \\
&= \sqrt{\frac{\Sigma_{y\mathbf{x}}\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{x}y}}{\Sigma yy}}.
\end{aligned}
$$

练习 4.4 假设 x, y, z 是三个随机变量, 其 Pearson 相关系数矩阵为 3×3 半正定矩阵

$$
R = \begin{pmatrix} 1 & \rho_{xy} & \rho_{xz} \\ \rho_{yx} & 1 & \rho_{yz} \\ \rho_{zx} & \rho_{zy} & 1 \end{pmatrix}.
$$

1. 证明
$$1 - \rho_{xy}^2 - \rho_{xz}^2 - \rho_{zy}^2 + 2\rho_{xy}\rho_{yz}\rho_{zx} \geq 0.$$
2. 假设 $\rho_{xy} = \rho_{xz} = \rho_{yz} = \rho$, 说明必定 $\rho \geq -1/2$。

解

1. 等价于原矩阵行列式大于等于 0.
2. 直接求解上问所得方程。

练习 4.5 假设 $n\times 1$ 随机向量 $\mathbf{x} = (x_1, ..., x_n)^\top$ 的诸分量之间的 Pearson 相关系数都是 ρ, 试证明 $1 \geq \rho \geq -1/(n-1)$。并求偏相关系数 $\rho_{x_1x_2\bullet x_3\cdots x_n}$。

解

1. 首先，随机变量之间相关系数是一个归一化的量，与他们的方差无关，所以不妨假设所有分量的方差都是 1，否则对每个随机变量除以它的标准差进行研究。由此可以得到协方差矩阵 Σ 等于相关系数矩阵：
$$
R_n = \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{pmatrix}.
$$
注意到 R 的对角元为 1, 非对角元均为 ρ, 因此可以写成
$$R = (1-\rho)I_n + \rho\mathbf{1}\mathbf{1}^T.$$
利用 **Matrix determinant lemma**, 或者用一些初等的线性代数技巧，可以求得
$$\det(R) = (1 + \rho\mathbf{1}^T(1-\rho)^{-1}I_n\mathbf{1})\det\left((1-\rho)I_n\right) = \left(1 + \frac{n\rho}{1-\rho}\right)(1-\rho)^n.$$
再利用 R 半正定，行列式非负，易得 $-\frac{1}{n-1} \leq \rho \leq 1$.
以上两种求行列式的方法都不容易，还可以从半正定本身的定义出发，由
$$\mathbf{1}^T R\mathbf{1} \geq 0, \text{where } \mathbf{1} = (1, 1, 1, \cdots, 1)^T$$
直接推出 $-\frac{1}{n-1} \leq \rho \leq 1$ 这个结果。

2. 首先给出求形如 R_n 的矩阵的逆的简便方法，即直接利用 **Sherman–Morrison formula**, 可得

$$
\begin{aligned}
R_n^{-1} &= ((1-\rho)I_n + \rho\mathbf{1}\mathbf{1}^T)^{-1} \\
&= \frac{1}{1-\rho}I_n - \frac{1}{(1-\rho)^2}\frac{\rho\mathbf{1}\mathbf{1}^T}{1+\rho/(1-\rho)\mathbf{1}^T\mathbf{1}} \\
&= \frac{1}{(1-\rho)\big(1+(n-1)\rho\big)}\begin{pmatrix} 1+(n-2)\rho & -\rho & \cdots & -\rho \\ -\rho & 1+(n-2)\rho & \cdots & -\rho \\ \vdots & \vdots & \ddots & \vdots \\ -\rho & -\rho & \cdots & 1+(n-2)\rho \end{pmatrix}
\end{aligned}
$$

关于求偏相关系数，由于课上只给出了 $\rho_{x_1,x_2\bullet x_3}$ 的显示表达式，没有给更多元，不建议使用递推的方法，而要将多元的情况看作向量形式，如直接将 $\boldsymbol{v} = (x_3, \cdots, x_n)$ 看成向量，用矩阵方法求向量版本的偏相关系数即可。令 $\boldsymbol{u} = (x_1, x_2)$, 则 $\boldsymbol{x} = (\boldsymbol{u}, \boldsymbol{v})$.

此时，

$$
\begin{aligned}
\Sigma_{\boldsymbol{uu}\bullet\boldsymbol{v}} &= \Sigma_{\boldsymbol{uu}} - \Sigma_{\boldsymbol{uv}}\Sigma_{\boldsymbol{vv}}^{-1}\Sigma_{\boldsymbol{vu}} \\
&= \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} - \begin{pmatrix} \rho & \cdots & \rho \\ \rho & \cdots & \rho \end{pmatrix} R_{n-2}^{-1} \begin{pmatrix} \rho & \rho \\ \vdots & \vdots \\ \rho & \rho \end{pmatrix} \qquad = \begin{pmatrix} Var(x_1^{\perp}) & cov(x_1^{\perp}, x_2^{\perp}) \\ cov(x_2^{\perp}, x_1^{\perp}) & Var(x_2^{\perp}) \end{pmatrix}.
\end{aligned}
$$

再利用偏相关系数的定义即可得到结果

$$
\rho_{x_1x_2\bullet x_3\cdots x_n} = \frac{cov(x_1^{\perp}, x_2^{\perp})}{\sqrt{Var(x_1^{\perp})Var(x_2^{\perp})}} = \frac{\rho}{1+(n-2)\rho}.
$$

最后，在已知 R^{-1} 的情况下，一种利用课上结论更简单的方法是，利用课件 *week04 P14* 命题 2 直接得到

$$
\rho_{12\bullet\boldsymbol{v}} = -\omega_{ij}/\sqrt{\omega_{11}\omega_{22}} = \frac{\rho}{1+(n-2)\rho},
$$

其中, $\Omega = (\omega_{ij}) = R^{-1}$.

4.3.2 补充内容

命题 4.1 (Matrix determinant lemma)

Suppose A is an invertible square matrix and u, v are column vectors. Then the matrix determinant lemma states that

$$
\det\left(\mathbf{A} + \mathbf{u}\mathbf{v}^{\mathsf{T}}\right) = \left(1 + \mathbf{v}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{u}\right)\det\left(\mathbf{A}\right).
$$

命题 4.2 (Sherman–Morrison formula)

Suppose $A \in \mathbb{R}^{n\times n}$ is an invertible square matrix and $u, v \in \mathbb{R}^n$ are column vectors. Then $A + uv^T$ is invertible iff $1 + v^TA^{-1}u \neq 0$. In this case,

$$
\left(A + uv^T\right)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}.
$$

Here, uv^T is the outer product of two vectors u and v.

以上两个公式更大的意义在于设计高效的算法。例如当计算样本的 $d \times d$ 维协方差矩阵 $\Sigma_n = \Sigma_{i=1}^n x_i x_i^T$ 的行列式或逆矩阵时，直接计算的复杂度相当高，但可以将原矩阵写为 $\Sigma_n = \Sigma_{n-1} + x_n x_n^T$, 这样在第 n 步便可以直接利用 $n-1$ 步的结果，再进行简单的矩阵乘法，而不必直接处理一个 $d \times d$ 维高阶矩阵。

4.4 HW 4

作业 4 链接

练习 4.1 假设单因子模型

$$\begin{cases} x = z + \epsilon_1 \\ y = z + \epsilon_2 \end{cases}$$

其中随机变量 $z, \epsilon_1, \epsilon_2$ 独立, $\mathrm{var}(z) = \sigma^2, \mathrm{var}(\epsilon_i) = \tau_i^2, i = 1, 2.$

1. 求三元随机向量 $(x, y, z)^\top$ 的方差-协方差矩阵。
2. 求 x, y 关于 z 的去相关化 $x^\perp, y^\perp$ 。
3. 求偏相关系数 $\rho_{xy\bullet z}$ 。

解

1.

$$\begin{aligned} \mathrm{var}(x) &= \mathrm{var}(z) + \mathrm{var}(\varepsilon_1) = \sigma^2 + \tau_1^2 \\ \mathrm{cov}(x, y) &= \mathrm{cov}(z + \varepsilon_1, z + \varepsilon_2) = \sigma^2 \\ \mathrm{cov}(x, z) &= \mathrm{cov}(z + \varepsilon_1, z) = \sigma^2 \end{aligned}$$

故 (x, y, z) 的协方差阵为 $\begin{pmatrix} \sigma^2 + \tau_1^2 & \sigma^2 & \sigma^2 \\ \sigma^2 & \sigma^2 + \tau_2^2 & \sigma^2 \\ \sigma^2 & \sigma^2 & \sigma^2 \end{pmatrix}$

2.

$$\begin{aligned} x^\perp &= x - \frac{\sigma^2}{\sigma^2} z = \varepsilon_1 \\ y^\perp &= y - \frac{\sigma^2}{\sigma^2} z = \varepsilon_2 \end{aligned}$$

3.

$$\rho_{xy\cdot z} = \frac{\mathrm{cov}\left(x^\perp, y^\perp\right)}{\sqrt{\mathrm{Var}\left(x^\perp\right) \mathrm{Var}\left(y^\perp\right)}} = 0$$

练习 4.2

假设三元随机向量 $(x, y, z)^\top$ 的相关系数矩阵如下

$$\Sigma = \begin{pmatrix} 1 & \rho_{xy} & \rho_{xz} \\ \rho_{yx} & 1 & \rho_{yz} \\ \rho_{zx} & \rho_{zy} & 1 \end{pmatrix}$$

1. 求 y 与 (x, z) 的决定系数。
2. 求控制 z 后, x, y 的偏决定系数。

解

1.

$$\begin{aligned} &\begin{pmatrix} \rho_{yx} & \rho_{yz} \end{pmatrix} \begin{pmatrix} 1 & \rho_{xz} \\ \rho_{zx} & 1 \end{pmatrix}^{-1} \begin{pmatrix} \rho_{xy} \\ \rho_{zy} \end{pmatrix}. \\ =& \begin{pmatrix} \rho_{yx} & \rho_{yz} \end{pmatrix} \frac{1}{1 - \rho_{xz}^2} \begin{pmatrix} 1 & -\rho_{xz} \\ -\rho_{xz} & 1 \end{pmatrix} \begin{pmatrix} \rho_{xy} \\ \rho_{zy} \end{pmatrix} \\ =& \frac{1}{1 - \rho_{xz}^2} \left(\rho_{xy} - \rho_{yz}\rho_{xz} - \rho_{xy}\rho_{xz} + \rho_{yz}\right) \begin{pmatrix} \rho_{xy} \\ \rho_{zy} \end{pmatrix} \\ =& \frac{1}{1 - \rho_{xz}^2} \left(\rho_{xy}^2 - 2\rho_{xy}\rho_{yz}\rho_{zx} + \rho_{yz}^2\right). \end{aligned}$$

2. 可以直接按照偏决定系数的定义计算。当然，两个随机变量（一维）的偏决定系数就是偏相关系数的平方，而偏相关系数 $\rho_{xy\bullet z}$ 的表达式已知。故

$$\phi_{xy\bullet z} = \rho^2_{xy\bullet z} = \frac{(\rho_{xy} - \rho_{xz}\rho_{yz})^2}{(1-\rho^2_{xz})(1-\rho^2_{yz})}.$$

练习 4.3 假设 $n \times 1$ 随机向量 $\mathbf{x} = (x_1, \ldots, x_n)^\top$ 的各分量方差都是 1，每两个分量的协方差（即相关系数）都是 $\rho > 0$, 求 x_1 与 $(x_2, \ldots, x_n)^\top$ 的决定系数。

解 $\boldsymbol{x} = (x_1, \ldots, x_n)$ 的协方差阵形如

$$R = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

令 $\boldsymbol{u} = (x_2, x_3, \ldots, x_n)^T$, 则 x_1 关于 u 的决定系数可以写为:

$$\phi_1 = \frac{\Sigma_{x_1\boldsymbol{u}}\Sigma^{-1}_{\boldsymbol{uu}}\Sigma_{\boldsymbol{u}x_1}}{\Sigma_{x_1x_1}} = \frac{\rho^2(n-1)}{(n-2)\rho+1},$$

便是我们要的答案。这里我们可以像上次作业一样计算逆矩阵

$$\Sigma^{-1}_{\boldsymbol{uu}} = \frac{1}{((n-2)\rho+1)(1-\rho)} \begin{pmatrix} (n-3)\rho+1 & -\rho & \cdots & -\rho \\ -\rho & (n-3)\rho+1 & \cdots & -\rho \\ \vdots & \vdots & \ddots & \vdots \\ -\rho & -\rho & \cdots & (n-3)\rho+1 \end{pmatrix}.$$

另外，有同学在作业中提供了不用求逆的巧妙方法，可以直接猜出以下式子的值:

$$\begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}^{-1}_{(n-1\times n-1)} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

假设

$$\begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix},$$

同乘协方差矩阵，得

$$\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix} \begin{pmatrix} x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix},$$

由对称性，容易看出解为

$$\begin{pmatrix} x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \frac{1}{(n-2)\rho+1} \\ \frac{1}{(n-2)\rho+1} \\ \vdots \\ \frac{1}{(n-2)\rho+1} \end{pmatrix}.$$

因此容易计算形如

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

的值，代入原式即可得到答案。

4.5 HW 5

作业 5 链接

练习 4.1 若 $\mathbf{x}_{m\times 1}, \mathbf{y}_{n\times 1}$ 为任意两个随机向量, $E(\mathbf{y} \mid \mathbf{x})$ 称为回归函数, 记 $\boldsymbol{\epsilon} = \mathbf{y} - E(\mathbf{y} \mid \mathbf{x})$.

(a) 证明 $\boldsymbol{\epsilon}$ 与 $\mathbf{x}$ 不相关 (因此, $\mathbf{y}$ 可分解为两个不相关的部分: $\mathbf{y} = E(\mathbf{y} \mid \mathbf{x}) + \boldsymbol{\epsilon}$).

(b) 利用 (a) 的结果证明 $\min_{f:R^m \to R^n} E\|\mathbf{y} - f(\mathbf{x})\|^2 = E\|\mathbf{y} - E(\mathbf{y} \mid \mathbf{x})\|^2$ （提示: 首先证明 $\hat{f}(\mathbf{x}) = E(\mathbf{y} \mid \mathbf{x})$ 极小化给定 $\mathbf{x}$ 条件下的条件期望 $E\left(\|\mathbf{y} - f(\mathbf{x})\|^2 \mid \mathbf{x}\right)$).

(c) 利用 (a) 的结果证明方差矩阵分解公式: $\operatorname{var}(\mathbf{y}) = \operatorname{var}[E(\mathbf{y} \mid \mathbf{x})] + E[\operatorname{var}(\mathbf{y} \mid \mathbf{x})]$ (提示: 只需证明 $\operatorname{var}(\boldsymbol{\epsilon}) = E[\operatorname{var}(\mathbf{y} \mid \mathbf{x})]$).

(d) 假设 $\mathbf{x}, \mathbf{y}, \mathbf{z}$ 是三个随机向量, 证明

$$\operatorname{cov}(\mathbf{x}, \mathbf{y}) = \operatorname{cov}[E(\mathbf{x} \mid \mathbf{z}), E(\mathbf{y} \mid \mathbf{z})] + E[\operatorname{cov}(\mathbf{x}, \mathbf{y} \mid \mathbf{z})].$$

证明

(a)

$$\begin{aligned}
&\operatorname{Cov}(\varepsilon, x) \\
=&\operatorname{Cov}(y - E(y \mid x), x) \\
=&E\left((y - E(y \mid x))x^\top\right) - E(y - E(y \mid x))Ex^\top \\
=&E\left((y - E(y \mid x))x^\top\right) \\
=&0 \quad \text{（根据条件期望定义）}
\end{aligned}$$

(b)

$$\begin{aligned}
&E\left(\|y - f(x)\|^2 \mid x\right) \\
=&E\left(\|y - E(y \mid x) + E(y \mid x) - f(x)\|^2 \mid x\right)
\end{aligned}$$

$$\begin{aligned}
&\text{记}\varepsilon = y - E(y \mid x) \\
&\text{由于}\varepsilon\text{与}x\text{ 不相关} \\
\Rightarrow&\operatorname{Cov}(\varepsilon, E(y \mid x) - f(x)) = 0 \\
\Rightarrow&E\left(\varepsilon(E(y \mid x) - f(x))^\top\right) = 0
\end{aligned}$$

$$\begin{aligned}
&E\left(\|y - f(x)\|^2 \mid x\right) \\
=&E\left(\|\varepsilon\|^2 \mid x\right) + E\left(\|E(y \mid x) - f(x)\|^2 \mid x\right) \\
=&E\left(\|\varepsilon\|^2 \mid x\right) + \|E(y \mid x) - f(x)\|^2 \\
\geqslant&E\left(\|\varepsilon\|^2 \mid x\right)
\end{aligned}$$

$$\begin{aligned}
&E\|y - f(x)\|^2 \\
=&E\left(E\left(\|y - f(x)\|^2 \mid x\right)\right) \\
\geqslant&E\left(E\left(\|\varepsilon\|^2 \mid x\right)\right) \\
=&E\|\varepsilon\|^2 \\
&\text{当且仅当}f(x) = E(y \mid x)
\end{aligned}$$

(c)

$$
\begin{aligned}
&\operatorname{Var}(\varepsilon)\\
=&E\left(\varepsilon\varepsilon^{\top}\right)-(E\varepsilon)(E\varepsilon)^{\top}\\
=&E\left(E\left(\varepsilon\varepsilon^{\top}\mid x\right)\right)\\
=&E((y-E|y|x))(y-E(y\mid x))^{\top}\mid x)\\
=&E\left(E\left(yy^{\top}+E(y\mid x)E(y\mid x)^{\top}-2yE(y\mid x)^{\top}\mid x\right)\right)\\
=&E\left(E\left(yy^{\top}\mid x\right)+E(y\mid x)E(y\mid x)^{\top}-2E(y\mid x)E(y\mid x)^{\top}\right)\\
=&E\left(E\left(yy^{\top}\mid x\right)-E(y\mid x)E(y\mid x)^{\top}\right)\\
=&E(\operatorname{Var}(y\mid x))
\end{aligned}
$$

(d)

$$
\begin{aligned}
&\operatorname{Cov}(E(x\mid z),E(y\mid z))\\
=&E\left(E(x\mid z)E(y\mid z)^{\top}\right)-ExEy^{\top}
\end{aligned}
$$

$$
\begin{aligned}
&E(\operatorname{Cov}(x,y\mid z))\\
=&E\left(E(xy^{\top}\mid z)-E(x\mid z)E(y\mid z)^{\top}\right)\\
=&Exy^{\top}-E\left(E(x\mid z)E(y\mid z)^{\top}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\operatorname{Cov}(E(x\mid z),E(y(z))+E(\operatorname{Cov}(x,y\mid z))\\
=&Exy^{\top}-ExEy^{\top}\\
=&\operatorname{Cov}(x,y)
\end{aligned}
$$

✍ **练习 4.2** 假设 y 是一元响应变量, $\mathbf{x}$ 是 q 元随机向量, z 是 $(p-q)\times 1$ 随机向量, 都是可观测变量。我们感兴趣的问题是研究 y 与 $\mathbf{x}$ 的关系, 而 $\mathbf{z}$ 是干扰因素 (既与 y 有关也与 $\mathbf{x}$ 有关), 需要加以控制。假设如下线性模型

$$y=a+\mathbf{b}^{\top}\mathbf{x}+\mathbf{c}^{\top}\mathbf{z}+\epsilon,\epsilon\sim\left(0,\sigma^2\right),\epsilon\perp\mathbf{x},\mathbf{z}(1)$$

问题是, 该模型是否正确地实现了目标 “在控制 $\mathbf{z}$ 的条件下, 研究 y 与 $\mathbf{x}$ 的关系” ? 为了考察模型 (1) 中加项 $\mathbf{c}^{\top}\mathbf{z}$ 的作用, 以及考察系数 $\mathbf{b}$ 的含义, 我们令

$$y^{\perp}=y-\Sigma_{y\mathbf{z}}\Sigma_{\mathbf{zz}}^{-1}\mathbf{z},\quad \mathbf{x}^{\perp}=\mathbf{x}-\Sigma_{\mathbf{xz}}\Sigma_{\mathbf{zz}}^{-1}\mathbf{z},$$

则可以证明模型 (1) 蕴含了

$$y^{\perp}=a+\mathbf{b}^{\top}\mathbf{x}^{\perp}+\epsilon,\quad \epsilon\sim\left(0,\sigma^2\right),\quad \epsilon\perp\mathbf{x}^{\perp}.(2)$$

这说明模型 (1) 中线性项 $\mathbf{c}^{\top}\mathbf{z}$ 确实起到了控制（或消除）$\mathbf{z}$ 影响的作用。进一步, 可以证明

$$\mathbf{b}=\Sigma_{\mathbf{xx}\bullet\mathbf{z}}^{-1}\Sigma_{\mathbf{x}y\bullet\mathbf{z}},(3)$$

特别地, 当 $q=1$ 时, $b\propto\rho_{xy\bullet z}$ 。结果 (4) 说明了模型 (2) 中 $\mathbf{x}$ 的系数 $\mathbf{b}$ 确实代表了偏相关系数, 干扰因素的影响通过在模型 (1) 添加 $\mathbf{c}^{\top}\mathbf{z}$ 得到了消除。

另一方面, 假设模型确实是正确模型, 但假如我们没有意识到 z 是一个潜在的干扰因素, 并假设模型

$$y=\alpha+\beta^{\top}\mathbf{x}+\delta,\delta\sim\left(0,\tau^2\right),\delta\perp\mathbf{x}(4)$$

也就是说, 我们将模型 (1) 中的 $\mathbf{c}^{\top}\mathbf{z}$ 一项归入了误差项:

$$\delta=\epsilon+\mathbf{c}^{\top}\mathbf{z}-E\left(\mathbf{c}^{\top}\mathbf{z}\right)$$

其中最后一项 $-E\left(\mathbf{c}^{\top}\mathbf{z}\right)$ 是为了保证 $E(\delta)=0$, 而 $\alpha=a+E\left(\mathbf{c}^{\top}\mathbf{z}\right),\beta=\mathbf{b}$ 。显然 δ 与 $\mathbf{x}$ 不独立, 那么基于

错误模型 (4), 由课件命题 2

$$\mathbf{b} = \beta = \Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{x}y}, (5)$$

与 (3) 相比, 这个表达没有控制 **z**, 因而是错误的表达。本题的任务是, 假设模型 (1) 是正确模型, 证明 (2) 和 (3)。

证明

(2)

$$\begin{aligned}
y^{\perp} &= y - \Sigma_{yz}\Sigma_{zz}^{-1}z \\
&= a + b^{\top}x + c^{\top}z + \varepsilon - \Sigma_{yz}\Sigma_{zz}^{-1}z \\
&= a + b^{\top}\left(x^{\perp} + \Sigma_{xz}\Sigma_{zz}^{-1}z\right) + c^{\top}z + \varepsilon - \Sigma_{yz}\Sigma_{zz}^{-1}z \\
&= a + b^{\top}x^{\perp} + \varepsilon + c'^{\top}z
\end{aligned}$$

$$\begin{aligned}
&\text{由于} y^{\perp}x^{\perp}, \varepsilon \text{与} z \text{ 不相关} \\
\Rightarrow & c' = 0, \quad \varepsilon \perp x^{\perp} \\
\Rightarrow & y^{\perp} = a + b^{\top}x^{\perp} + \varepsilon
\end{aligned}$$

(3)

$$\varepsilon = y^{\perp} - a - b^{\top}x^{\perp} \text{与} x^{\perp} \text{ 不相关}$$

$$\operatorname{Cov}\left(\varepsilon, x^{\perp}\right) = 0$$

$$\Sigma_{yx\cdot z} - b^{\top}\Sigma_{xx\cdot z} = 0$$

$$b = \Sigma_{xx\cdot z}^{-1}\Sigma_{xy\cdot z}$$

4.6 HW 6

作业 6 链接

注 在做回归分析作业，考虑随机性时可以忽略样本自变量 x 带来的随机性，计算期望、方差时均可以看作是固定 x 以后的条件期望、条件方差。这会让许多概念变得更清楚。

注 事实上，在很多时候，x 的随机性也可以考虑到模型中。这时我们通常的做法是先固定 x 求出条件分布、条件期望等等，再考虑 x 的随机性对于结果的影响。例如讲义 *week 06 P20* 的几个结论，在不考虑 x 带来的随机性，即都看作对 x 取条件的版本，那证明都是比较容易的。**其实从这里也可以更进一步推出，当 x 具有随机性时对应的结论。(思考)**

✍ **练习 4.1** 简单随机样本 $y_1, \ldots, y_n$ iid $\sim \left(\mu, \sigma^2\right)$ 是最简单 (不含自变量) 的线性模型, 这是因为该模型可写为:

$$y_i = \mu + \epsilon_i, i = 1, \ldots, n, \quad \epsilon_i, i = 1, \ldots, n \text{ iid } \sim \left(0, \sigma^2\right)$$

其中 μ 是未知参数。使得误差平方和 $\sum_{i=1}^n \left(y_i - \mu\right)^2$ 最小的 μ 称为最小二乘估计, 记为 $\hat{\mu}$ 。求 $\hat{\mu}$ 及其方差。y_i 的拟合值 $\hat{y}_i$ 和残差 e_i 应如何定义?

注 建议回顾一下拟合值和残差的定义。拟合值是根据模型和数据可以给出的数值，而扰动项 ε_i 是无法观测的。能利用的数据只有 $y_i, i = 1, \ldots, n$.

解 误差平方和对 μ 求导等于 0, 解得 $\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$. $var(\bar{y}) = \sigma^2/n$.

拟合值为 $\hat{\mu} = \bar{y}$, 残差为 $y_i - \bar{y}$.

✍ **练习 4.2** 假设随机样本数据 $\left(x_i, y_i\right), i = 1, \ldots, n$ 满足下述过原点的线性回归模型 (无截距项)

$$y_i = bx_i + \epsilon_i, E\left(\epsilon_i\right) = 0, \operatorname{var}\left(\epsilon_i\right) = \sigma^2, \text{且} x_i \text{ 与} \epsilon_i \text{ 独立}, i = 1, \ldots, n.$$

通过极小化误差平方和 $\sum_{i=1}^n \left(y_i - bx_i\right)^2$, 求未知参数 b 的 LS 估计 $\widehat{b}$, 并求其方差。

解 同上题。误差平方和对 b 等于 0, 解得 $\hat{b} = \sum_{i=1}^n x_i y_i / \sum_{i=1}^n x_i^2 . var(\hat{b}) = \sigma^2 / \sum_{i=1}^n x_i^2$.

✍ **练习 4.3** 证明给定 $\mathbf{x} = \left(x_1, \ldots, x_n\right)^\top$ 条件下

$$\operatorname{var}(\hat{a}) = \sigma^2/n + \bar{x}^2\sigma^2/s_{xx}, \quad \operatorname{var}(\hat{b}) = \sigma^2/s_{xx}, \quad \operatorname{cov}(\hat{a}, \hat{b}) = -\bar{x}\sigma^2/s_{xx}$$

何时 $\hat{a}, \hat{b}$ 不相关?

证明

$$\begin{aligned}
var(\hat{b}) &= var\left(\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \\
&= var\left(\frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x})^2 var(y_i)}{s_{xx}^2} \\
&= \sigma^2/s_{xx}, \\
var(\hat{a}) &= var(\bar{y} - \hat{b}\bar{x}) \\
&= var\left(\frac{1}{n}\sum_{i=1}^n y_i - \hat{b}\bar{x}\right) \\
&= var\left(\frac{1}{n}\sum_{i=1}^n y_i - \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\bar{x}\right) \\
&= var\left(\sum_{i=1}^n \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{s_{xx}}\right) y_i\right) \\
&= \sum_{i=1}^n \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{s_{xx}}\right)^2 var(y_i) \\
&= \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{x}^2}{s_{xx}},
\end{aligned}$$

$$
\begin{aligned}
cov(\hat{a},\hat{b}) &= cov(\bar{y}-\hat{b}\bar{x},\hat{b}) \\
&= cov\left(\bar{y},\frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)-\bar{x}var(\hat{b}) \\
&= cov\left(\frac{1}{n}\sum_{i=1}^{n}y_i,\frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)-\bar{x}var(\hat{b}) \\
&= \sum_{i=1}^{n}\frac{1}{n}\frac{(x_i-\bar{x})}{s_{xx}}var(y_i)-\bar{x}var(\hat{b}) \\
&= 0-\frac{\bar{x}\sigma^2}{s_{xx}}=-\frac{\bar{x}\sigma^2}{s_{xx}}.
\end{aligned}
$$

$\hat{a},\hat{b}$ 不相关当且仅当 $cov(\hat{a},\hat{b})=0$, 即 $\bar{x}=0$ 或 $\sigma=0$.

练习 4.4 假设 $x_i=0$ 或 $1, i=1,\ldots,n$, 记 $m=\sum x_i$ 为 1 的个数。当 m 等于或近似等于 $n/2$ 时, 称为是均衡设计 (balanced design), 否则是不均衡的 (unbalanced)。基于 b 的 LS 估计的方差, 讨论均衡设计的优越性。

解 $var(\hat{b})=\sigma^2/s_{xx}$. 均衡设计使 LS 估计参数 $\hat{b}$ 的方差达到最小。

练习 4.5 定义最小二乘得到的残差 $e_i=y_i-\hat{y}_i=y_i-\hat{a}-\hat{b}x_i$, 残差平方和定义为 $RSS=\sum_{i=1}^{n}e_i^2$ 。证明 $e_i=(\epsilon_i-\bar{\epsilon})-(x_i-\bar{x})\,s_{x\epsilon}/s_{xx}$, 并利用该表达证明

$$RSS=s_{yy}-s_{xy}^2/s_{xx}=s_{\epsilon\epsilon}-s_{x\epsilon}^2/s_{xx}.$$

证明 证明参考课件 week 06 P16 等.

练习 4.6 定义点 (x_i,y_i) 与 $(\bar{x},\bar{y})$ 决定的直线的斜率为 $k_i=\frac{y_i-\bar{y}}{x_i-\bar{x}}$, 当 $x_i=\bar{x}$ 时定义 $k_i=0, i=1,\ldots,n$ 。则 b 的最小二乘估计 $\hat{b}=s_{xy}/s_{xx}$ 可表示成所有斜率 $k_i, i=1,\ldots,n$ 的加权和

$$\hat{b}=\frac{1}{s_{xx}}\sum(x_i-\bar{x})(y_i-\bar{y})=\frac{1}{s_{xx}}\sum(x_i-\bar{x})^2\left(\frac{y_i-\bar{y}}{x_i-\bar{x}}\right)=\sum w_{0i}k_i$$

其中权重 $w_{0i}=(x_i-\bar{x})^2/s_{xx}, \sum w_{0i}=1$ 。对任何序列 $w_1,\ldots,w_n$ (只依赖于 $x_1,\ldots,x_n$) , $w_i\geq 0, \sum w_i=1$, 定义 b 的一个估计

$$\tilde{b}=\sum w_ik_i=\sum w_i\left(\frac{y_i-\bar{y}}{x_i-\bar{x}}\right).$$

(a) 证明 $E(\tilde{b})=b$.

(b) 求 $\mathrm{var}(\tilde{b}\mid\mathbf{x})$.

(c) 证明 $\mathrm{var}(\tilde{b}\mid\mathbf{x})\geq\sigma^2/s_{xx}=\mathrm{var}(\hat{b}\mid\mathbf{x})$.

证明

(a)

$$
\begin{aligned}
E(\tilde{b}) &= E(E(\tilde{b}\mid x)) \\
&= E\left(\sum_{i=1}^{n}\frac{w_i}{x_i-\bar{x}}E\left(y_i-\bar{y}\mid x\right)\right) \\
&= E\left(\sum_{i=1}^{n}\frac{w_i}{x_i-\bar{x}}\left(b\left(x_i-\bar{x}\right)\right)\right) \\
&= bE\left(\sum_{i=1}^{n}w_i\right) \\
&= b
\end{aligned}
$$

(b) 注意到

$$\sum_{i=1}^{n}a_i\left(b_i-\bar{b}\right)=\sum_{i=1}^{n}\left(a_i-\bar{a}\right)b_i$$

$$
\begin{aligned}
\operatorname{Var}(\tilde{b} \mid x) &= \operatorname{Var}\left(\sum_{i=1}^{n} \frac{w_i}{x_i - \bar{x}} (y_i - \bar{y}) \mid x\right) \\
&= \operatorname{Var}\left(\sum_{i=1}^{n} c_i (y_i - \bar{y}) \mid x\right) \qquad 记 c_i = \frac{w_i}{x_i - \bar{x}} \\
&= \operatorname{Var}\left(\sum_{i=1}^{n} (c_i - \bar{c}) y_i \mid x\right) \\
&= \sum_{i=1}^{n} (c_i - \bar{c})^2 \sigma^2 \\
&= \left(\sum_{i=1}^{n} c_i^2 - n\bar{c}^2\right) \sigma^2 \\
&= \left(\sum_{i=1}^{n} \frac{w_i^2}{(x_i - \bar{x})^2} - \frac{1}{n}\left(\sum_{i=1}^{n} \frac{w_i}{x_i - \bar{x}}\right)^2\right) \sigma^2
\end{aligned}
$$

(c)

$$
\operatorname{Var}(\tilde{b} \mid x) = \sum_{i=1}^{n} (c_i - \bar{c})^2 \sigma^2
$$

由 Cauchy - Schwarz 不等式

$$
\sum_{i=1}^{n} (c_i - \bar{c})^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \geqslant \left(\sum_{i=1}^{n} (c_i - \bar{c})(x_i - \bar{x})\right)^2
$$

$$
\sum_{i=1}^{n} (c_i - \bar{c})^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{\operatorname{Var}(\tilde{b} \mid x) s_{xx}}{\sigma^2}
$$

$$
\begin{aligned}
\sum_{i=1}^{n} (c_i - \bar{c})(x_i - \bar{x}) &= \sum_{i=1}^{n} c_i (x_i - \bar{x}) \\
&= \sum_{i=1}^{n} \omega_i \\
&= 1
\end{aligned}
$$

$$
\operatorname{Var}(\tilde{b} \mid x) \geqslant \frac{\sigma^2}{s_{xx}}
$$

注 本题问题多出在方差的计算上，公式

$$
\operatorname{Var}(X_1 + \cdots + X_n) = \operatorname{Var}(X_1) + \cdots + \operatorname{Var}(X_n)
$$

只能用于独立的随机变量。而形如 $y_i - \bar{y}$ $(i = 1, \ldots, n)$ 的随机变量列并不是相互独立的，应该拆成独立的部分计算。

注 另外，本题实质上是简单形式的 *Gauss-Markov* 定理。该定理可以表述为：若 $\tilde{b} = \Sigma_{i=1}^{n} w_i y_i$, $(\{w_i\}_{i=1}^{n}$ 只与 x 有关) 是 b 的无偏估计，首先我们知道通过选取适当的 w $\tilde{b}$ 可以取为最小二乘估计 $\hat{b}$，并且最小二乘估计 $\hat{b}$ 是对应不同 w 的无偏估计 $\tilde{b}$ 中，方差最小的一个。

4.7 HW 7

作业 7 链接

下列各题都基于简单线性回归模型：假设独立样本 $(x_i, y_i), i = 1, \ldots, n$ 满足下述模型

$$y_i = a + bx_i + \epsilon_i, \epsilon_i \sim \left(0, \sigma^2\right), \text{且} x_i \text{ 与} \epsilon_i \text{ 独立}, i = 1, \ldots, n.$$

未知参数 a, b, σ^2 的 LS 估计分别记为 $\hat{a}, \hat{b}, \hat{\sigma}^2$. 记 $\mathbf{x} = (x_1, \ldots, x_n)^\top$.

练习 4.1 对于简单线性模型, 证明 $r_{\hat{y}y} = |r_{xy}|$ (前者为拟合值与响应变量的相关系数)。

证明 由

$$\hat{y}_i - \bar{\hat{y}} = \left(\hat{a} + \hat{b}x_i\right) - (\hat{a} + \hat{b}\bar{x}) = \hat{b}\left(x_i - \bar{x}\right),$$

得

$$\begin{aligned}
r_{\hat{y}y} &= \frac{\sum_{i=1}^n \left(\hat{y}_i - \bar{\hat{y}}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^n \left(\hat{y}_i - \bar{\hat{y}}\right)^2}\sqrt{\sum_{i=1}^n \left(y_i - \bar{y}\right)^2}} \\
&= \frac{\hat{b}\sum_{i=1}^n \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\hat{b}^2\sum_{i=1}^n \left(x_i - \bar{x}\right)^2}\sqrt{\sum_{i=1}^n \left(y_i - \bar{y}\right)^2}} \\
&= \operatorname{Sgn}(\hat{b}) \cdot \frac{\sum_{i=1}^n \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^n \left(x_i - \bar{x}\right)^2}\sqrt{\sum_{i=1}^n \left(y_i - \bar{y}\right)^2}} \\
&= \operatorname{Sgn}\left(r_{xy}\right) \cdot r_{xy} = \left|r_{xy}\right|.
\end{aligned}$$

练习 4.2 定义最小二乘得到的残差 $e_i = y_i - \hat{y}_i = y_i - \hat{a} - \hat{b}x_i$, 证明

$$E\left(e_i\right) = 0, \quad \operatorname{var}\left(e_i \mid \mathbf{x}\right) = \left(1 - \frac{1}{n} - \frac{\left(x_i - \bar{x}\right)^2}{s_{xx}}\right)\sigma^2.$$

由此证明 $\hat{\sigma}^2$ 是 σ^2 的无偏估计。

证明 由 HW6 第三题：

$$\operatorname{var}(\hat{a}) = \sigma^2/n + \bar{x}^2\sigma^2/s_{xx}, \quad \operatorname{var}(\hat{b}) = \sigma^2/s_{xx}, \quad \operatorname{cov}(\hat{a}, \hat{b}) = -\bar{x}\sigma^2/s_{xx}.$$

$$\begin{aligned}
cov(y_i, \hat{b}) &= cov\left(y_i, \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{s_{xx}}\right) \\
&= (x_i - \bar{x})/s_{xx} var(y_i) \\
&= (x_i - \bar{x})/s_{xx}\sigma^2. \\
cov(y_i, \hat{a}) &= cov(y_i, \bar{y} - \hat{b}\bar{x}) \\
&= cov(y_i, \bar{y}) - \bar{x} cov(y_i, \hat{b}) \\
&= \frac{1}{n}\sigma^2 - \bar{x}(x_i - \bar{x})/s_{xx}\sigma^2.
\end{aligned}$$

于是

$$
\begin{aligned}
var(e_i) &= var(y_i - \hat{a} - \hat{b}x_i) \\
&= cov(y_i - \hat{a} - \hat{b}x_i, y_i - \hat{a} - \hat{b}x_i) \\
&= var(y_i) + var(\hat{a}) + x_i^2 var(\hat{b}) - 2cov(y_i, \hat{a}) - 2x_i cov(y_i, \hat{b}) + 2x_i cov(\hat{a}, \hat{b}) \\
&= \sigma^2 + \sigma^2/n + \bar{x}^2\sigma^2/s_{xx} + x_i^2/s_{xx}\sigma^2 - 2\frac{1}{n}\sigma^2 + 2\bar{x}(x_i - \bar{x})/s_{xx}\sigma^2 - 2x_i(x_i - \bar{x})/s_{xx}\sigma^2 - 2x_i\bar{x}\sigma^2/s_{xx} \\
&= \left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{s_{xx}}\right)\sigma^2.
\end{aligned}
$$

$$
\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2,
$$

故

$$
E\hat{\sigma}^2 = \frac{1}{n-2}E\sum_{i=1}^{n} e_i^2 = \frac{1}{n-2}\sum_{i=1}^{n} var(e_i) = \sigma^2.
$$

练习 4.3 证明第六讲命题 5 。

证明 命题和证明参见 week 06 P20-21.

练习 4.4 对任一给定的 $x_0 \in R$, 均值函数或回归函数 $m(x_0) = E(y \mid x = x_0) = a + bx_0$ 的 LS 估计为

$$
\hat{m}(x_0) = \hat{a} + \hat{b}x_0
$$

其中 $\hat{a}, \hat{b}$ 是 a, b 的 LS 估计。

(a) 证明 $E(\hat{m}(x_0)) = m(x_0)$, $\operatorname{var}(\hat{m}(x_0) \mid \mathbf{x}) = \frac{\sigma^2}{n} + \frac{(x_0-\bar{x})^2\sigma^2}{s_{xx}}$.

(b) 假设误差 $\epsilon_1, \ldots, \epsilon_n$ iid $\sim N(0, \sigma^2)$, 证明 $\hat{m}(x_0)$ 与 $\hat{\sigma}^2$ 独立, 且

$$
\frac{\hat{m}(x_0) - m(x_0)}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{s_{xx}}}} \sim t_{n-2}
$$

证明

- $$
\begin{aligned}
E(\hat{m}(x_0)) &= E\hat{a} + E\hat{b}x_0 = a + bx_0 = m(x_0) \\
var(\hat{m}(x_0)) &= var(\hat{a} + \hat{b}x_0) = var(\hat{a}) + var(\hat{b})x_0^2 + 2x_0 cov(\hat{a}, \hat{b}) = \sigma^2(1/n + (x_0 - \bar{x})^2/s_{xx}).
\end{aligned}
$$

- 接下来的证明类似上题, 只是在记号上稍作改变。
  给定 x_0 和 $\boldsymbol{x}$，由于 $\hat{\sigma}^2$ 与 $\hat{a}$ 和 $\hat{b}$ 独立（我们可以从上一题中学到这一点），我们有

$$
\hat{m}(x_0) = \hat{a} + \hat{b}x_0 = g(\hat{a}, \hat{b}) \tag{4.1}
$$

  与 $\hat{\sigma^2}$ 无关。此外，请注意

$$
\hat{m}(x_0) - m(x_0) = \sum_{i=1}^{n}\left(\frac{(x_i - \bar{x})(x_0 - \bar{x})}{s_{xx}} + \frac{1}{n}\right)\varepsilon_i. \tag{4.2}
$$

  这给出

$$
\frac{\hat{m}(x_0) - m(x_0)}{\sigma\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{s_{xx}}}} = \frac{\sum_{i=1}^{n}\left(\frac{(x_i-\bar{x})(x_0-\bar{x})}{s_{xx}} + \frac{1}{n}\right)\varepsilon_i}{\sqrt{\sum_{i=1}^{n}\left(\frac{(x_i-\bar{x})(x_0-\bar{x})}{s_{xx}} + \frac{1}{n}\right)^2\sigma^2}} \sim \mathcal{N}(0,1). \tag{4.3}
$$

  因此，

$$
\frac{\hat{m}(x_0) - m(x_0)}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{s_{xx}}}} = \frac{\hat{m}(x_0) - m(x_0)}{\sigma\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{s_{xx}}}} \Big/ \sqrt{\frac{(n-2)\hat{\sigma}^2}{(n-2)\sigma^2}} \sim t_{n-2}. \tag{4.4}
$$

练习 4.5

简单线性模型的斜率估计 $\hat{b}=s_{xy}/s_{xx}=\sum(x_i-\bar{x})\,y_i/s_{xx}=\sum c_{0i}y_i$ 是 $y_1,\ldots,y_n$ 的线性组合, 其中 $c_{0i}=(x_i-\bar{x})/s_{xx}$ 。假设 $\widetilde{b}=\sum c_iy_i$ 是 b 的任一线性无偏估计, 其中 $c_1,\ldots,c_n$ 只与 $\mathbf{x}=(x_1,\ldots,x_n)^\top$ 有关, 与 y_i' 's 无关。

(a) 由 $\widetilde{b}$ 的无偏性, 证明 $c_1,\ldots,c_n$ 满足约束 $\sum c_i=0,\sum c_ix_i=1$ 。

(b) (Gauss-Markov 定理的特殊情况) 证明 $\mathrm{var}(\widetilde{b}\mid\mathbf{x})\geq\sigma^2/s_{xx}=\mathrm{var}(\hat{b}\mid\mathbf{x})$.

证明 (a) 由于对于任意 $b\in\mathbb{R}$ 都成立

$$b=\mathbb{E}\tilde{b}=\sum_{i=1}^n c_i\mathbb{E}y_i=\sum_{i=1}^n c_i\left(a+bx_i\right)=a+b\sum_{i=1}^n c_ix_i$$

我们有

$$\sum_{i=1}^n c_i=0,\quad \sum_{i=1}^n c_ix_i=1$$

(b) 注意到

$$\tilde{b}-b=\sum_{i=1}^n c_i\left(y_i-a-bx_i\right)=\sum_{i=1}^n c_i\varepsilon_i$$

由此得到

$$\mathrm{Var}(\tilde{b}\mid x)=\sigma^2\sum_{i=1}^n c_i^2$$

应用 Cauchy-Schwarz 不等式，我们有

$$\left(\sum_{i=1}^n c_i^2\right)\left(\sum_{i=1}^n(x_i-\bar{x})^2\right)\geq\left(\sum_{i=1}^n c_i\left(x_i-\bar{x}\right)\right)^2=\left(\sum_{i=1}^n c_ix_i\right)^2=1$$

因此

$$\mathrm{Var}(\tilde{b}\mid x)\geq\frac{\sigma^2}{s_{xx}}=\mathrm{Var}(\hat{b}\mid x)$$

以下两题不是编程题，要求不借助 R 语言，只利用题目中的数据来算出其他的量。

练习 4.6 2000 年联合国的关于 193 个国家或地区的人口统计数据, 包括每个国家 (或地区) 的女性人均生育数目 (Fertility) 和人均国民生产总值 (PPgdp, 单位: 千美元)。部分数据如下。

	Fertility	PPgdp
Afghanistan	6.80	98
Albania	2.28	1317
Algeria	2.80	1784
Angola	7.20	739
Argentina	2.44	7163
Armenia	1.15	687
Australia	1.70	18788
—.		
(完整数据集参见 R package alr4 中的 UN1)		

考虑线性模型

$$\text{Fertility}\ =a+b\times\text{PPgdp}+\epsilon,\quad \epsilon\sim\left(0,\sigma^2\right)$$

```
Call: lm(formula = Fertility ~ PPgdp, data = UN1)

Coefficients:
              Estimate   Std. Error   t value     Pr(>|t|)
(Intercept)    3.733        ①         28.040    < 2e-16 ***
PPgdp         -0.085      0.012         ②       1.15e-11 ***
Residual standard error: 1.526 on  ③  degrees of freedom
Multiple R-squared:  ④
```

下面是 R 软件的部分输出结果:

(a) 填写 (1)-(4) 处的数字。

(b) 计算 Fertility 和 PPgdp 的样本方差和样本相关系数。

(c) 已知所有 193 个国家或地区的 PPgdp 的平均值为 6408 美元, 求全世界（即 193 个国家或地区) 的 Fertility 的平均值。

(d) 试解释 PPgdp 的回归系数估计值 -0.085 的含义。

解

- ①$std.error = \frac{Estimate}{t-value} = 0.133$,
  ②$t - value = \frac{Estimate}{std.error} = -7.083$,
  ③自由度 $= n - p = 191$,
  ④$t = \sqrt{n-2}\frac{r}{1-r^2} \Rightarrow R^2 = \frac{t^2/(n-2)}{1+t^2/(n-2)} = \frac{7.083^2/191}{1+7.083^2/191} = 0.208$.
- 由于

$$\mathrm{se}(\hat{b}) = \sqrt{\frac{\hat{\sigma^2}}{s_{xx}}}, \quad \hat{b} = \frac{s_{xy}}{s_{xx}}, \quad r = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}},$$

  我们有

$$s_{xx} = \frac{\hat{\sigma^2}}{\mathrm{se}(\hat{b})^2} = 16171.36, \quad s_{xy} = \hat{b}s_{xx} = -1374.566,$$
$$s_{yy} = \frac{s_{xy}^2}{s_{xx} \cdot r^2} = 561.6152, \quad r = -\sqrt{r^2} = -0.456$$

  由此得到

$$\mathcal{S}_x = \frac{1}{n-1}s_{xx} = 84.226, \quad \mathcal{S}_y = \frac{1}{n-1}s_{yy} = 2.925, \quad r_{xy} = r = -0.456$$

- 由于 $\bar{y} = \hat{a} + \hat{b}\bar{x}$，我们有

$$\overline{\text{Fertility}} = \hat{a} + \hat{b}\,\overline{\text{Education}} = 3.188$$

- 若一个国家比另一个国家的 PPgdp 高 1000 美元，则前者比后者的女性人均剩余数目大约低 -0.085 个.

注 关联而非因果！这是观察数据的回归结果。

练习 4.7

老忠实 (或老实) 喷泉 (Old faithful geyser) 是美国黄石公园的一个间歇式热喷泉。除了每天零点到清晨 6 点之间, 1980 年 10 月份的所有喷水持续时间 (Duration, 单位: 秒) 以及到下一次喷发的间隔时间 (Interval, 单位: 分钟) 被记录了下来, 共有 270 条记录 (数据集 alr4: oldfaith), 例如前 5 条记录如下:

Duration	Interval
216	79
108	54
200	74
137	62
272	85
...	

其中第一次喷水持续 216 秒, 其后经过 79 分钟再次喷水并持续了 108 秒, 等等。Duration (y) 和 Interval (x) 的平均值分别是 209.9 秒和 71.1 分钟, (Duration,Interval) 的样本协方差矩阵为

$$S=\begin{pmatrix} S_{yy} & S_{yx} \\ S_{xy} & S_{xx} \end{pmatrix}=\frac{1}{n-1}\begin{pmatrix} s_{yy} & s_{yx} \\ s_{xy} & s_{xx} \end{pmatrix}=\begin{pmatrix} 4677.5 & 827.3 \\ 827.3 & 182.2 \end{pmatrix}$$

注意区分其中的记号, 其中小写 $s_{ab}=\sum(a_i-\bar{a})\left(b_i-\bar{b}\right)$, 大写 $S_{ab}=s_{ab}/(n-1)$ 为样本协方差或方差。假设如下线性模型

$$\text{Duration }=a+b\times\text{ Interval }+\epsilon,\epsilon\sim\left(0,\sigma^2\right),$$

(a) (原题有误) 试求 LS 估计 $\hat{a},\hat{b}$ 。如果等待了很久，你预期下次喷水时间较长还是较短?

(b) 求 LS 估计 $\hat{\sigma}^2$ 及其标准差, 以及 $H_0:b=0$ 的 t 检验统计量;

(c) 求回归方程的决定系数 R^2;

(d) (原题有误) 如果某人已经等待了一小时，他预测下次喷水不少于多长时间?

解

(a) 由于 $\hat{b}=\frac{s_{xy}}{s_{xx}}$ 且 $\hat{a}=\bar{y}-\hat{b}\bar{x}$，我们有 $\hat{b}=4.541$ 和 $\hat{a}=-112.938$。如果这个模型工作良好，那么在等待了很久的情况下，下一次喷水时间较长。(因为 $\hat{b}>0$).

(b) 由于 $\hat{\sigma^2}=\frac{1}{n-2}RSS=\frac{1}{n-2}\left(s_{yy}-\frac{s_{xy}^2}{s_{xx}}\right)=924.90$。此外，我们知道 $H_0:b=0$ 的 t 统计量是

$$t=\frac{\hat{b}}{se(\hat{b})}=\frac{\sqrt{s_{xx}}\hat{b}}{\hat{\sigma}}=33.061$$

(c) 有

$$R^2=\frac{t^2}{t^2+(n-2)}=0.803$$

(d) 从线性模型中我们可以得知

$$\text{Duration}=\hat{a}+\hat{b}\times\text{Interval}=159.522\quad(\text{秒})$$

4.8 HW 8

作业 8 链接

✍ **练习 4.1** 设 A 是一个 $n \times m$ 实数矩阵, $\mathbf{x} \in R^m$, 证明 $A\mathbf{x} = 0 \Leftrightarrow A^\top A\mathbf{x} = 0$.

证明

($\Rightarrow$) 显然

($\Leftarrow$)

$$\begin{aligned} A^\top Ax &= 0 \\ x^\top A^\top Ax &= 0 \\ (Ax)^\top (Ax) &= 0 \\ Ax &= 0 \end{aligned}$$

✍ **练习 4.2** 若 $A_1^\top A_1 B A_2^\top A_2 = 0$, 则 $A_1 B A_2^\top = 0$, 其中 A_1, A_2, B 都是矩阵。

证明

由于 $Ax = 0 \Leftrightarrow A^\top Ax = 0$

$$\begin{aligned} & A_1^\top A_1 B A_2^\top A_2 = 0 \\ \Rightarrow & A_1 B A_2^\top A_2 = 0 \\ \Rightarrow & A_1 B A_2^\top = 0 \end{aligned}$$

✍ **练习 4.3** 假设 $A_{n\times m}$ 有奇异值分解 $A = UDV^\top$, 其中 U, V, D 分别是 $n \times r, m \times r, r \times r$ 矩阵 ($r = \operatorname{rank}(A)$), 满足 $U^\top U = V^\top V = I_r, D = \operatorname{diag}\left(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r}\right)$ (可逆), 证明 $C(U) = C(A), C(V) = C\left(A^\top\right)$ 。

证明

($C(U) \supseteq C(A)$) 显然

($C(U) \subseteq C(A)$)

$$\begin{aligned} \forall x, \quad Ux &= UDV^\top VD^{-1}x \\ &= A\left(VD^{-1}x\right) \in C(A) \end{aligned}$$

✍ **练习 4.4** 假设 A 是一个实数矩阵, $P_A = A\left(A^\top A\right)^- A^\top$ 为对应的投影阵, 证明 $C\left(P_A\right) = C(A)$ 。

证明

($C(P_A) \subseteq C(A)$) 显然

($C(P_A) \supseteq C(A)$)

$$\forall x, \quad Ax = P_A Ax \in C\left(P_A\right)$$

✍ **练习 4.5** 假设矩阵 $A_{n\times m} = B_{n\times r} C_{r\times m}$, 假设 C 是行满秩的, 证明 $C(A) = C(B), P_A = P_B$ 。

证明

($C(A) \subseteq C(B)$) 显然

($C(A) \supseteq C(B)$) 由于 C 行满秩, $\forall x \in \mathbb{R}^{r\times 1}$, 存在 $y \in \mathbb{R}^{m\times 1}$ 满足 $x = Cy$

$$Bx = BCy = Ay \in C(A)$$

✍ **练习 4.6** 按列划分 $A_{n\times m} = (A_1, A_2)$, 假设 $A_1^\top A_2 = 0$, 证明 $P_A = P_{A_1} + P_{A_2}$ 。

证明

$$
\begin{aligned}
P_A &= A\left(A^\top A\right)^- A^\top \\
&= \begin{pmatrix} A_1 & A_2 \end{pmatrix}\left(\begin{pmatrix} A_1^\top \\ A_2^\top \end{pmatrix}\begin{pmatrix} A_1 & A_2 \end{pmatrix}\right)^- \begin{pmatrix} A_1^\top \\ A_2^\top \end{pmatrix} \\
&= \begin{pmatrix} A_1 & A_2 \end{pmatrix}\begin{pmatrix} A_1^\top A_1 & 0 \\ 0 & A_2^\top A_2 \end{pmatrix}^- \begin{pmatrix} A_1^\top \\ A_2^\top \end{pmatrix} \\
&= \begin{pmatrix} A_1 & A_2 \end{pmatrix}\begin{pmatrix} \left(A_1^\top A_1\right)^- & 0 \\ 0 & \left(A_2^\top A_2\right)^- \end{pmatrix}\begin{pmatrix} A_1^\top \\ A_1 \end{pmatrix} \\
&= A_1\left(A_1^\top A_1\right)^- A_1^\top + A_2\left(A_2^\top A_2\right)^- A_2^\top \\
&= P_{A_1} + P_{A_2}
\end{aligned}
$$

✍ **练习 4.7** 按列划分 $A_{n\times m} = (A_1, A_2)$, 令 $A_2^\perp = A_2 - P_{A_1}A_2$, 证明 $P_A = P_{A_1} + P_{A_2}$.

证明

$$
\begin{aligned}
(A_2^\perp)^\top A_1 &= (A_2 - P_{A_1}A_2)^\top A_1 \\
&= (A_2^\top - A_2^\top P_{A_1})A_1 \\
&= A_2^\top A_1 - A_2^\top P_{A_1}A_1 \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
\tilde{A} &= \begin{pmatrix} A_1 & A_2^\perp \end{pmatrix} \\
&= \begin{pmatrix} A_1 & A_2 \end{pmatrix}\begin{pmatrix} I & -\left(A_1^\top A_1\right)A_1^\top A_2 \\ 0 & 1 \end{pmatrix} \\
&\begin{pmatrix} I & -\left(A_1^\top A_1\right)^- A_1^\top A_2 \\ 0 & I \end{pmatrix} \text{行满秩}
\end{aligned}
$$

$$
\text{由 5. 可知} \quad P_A = P_{\tilde{A}} = P_{A_1} + P_{A_2^\perp}
$$

✍ **练习 4.8** (选做) 通常, 逆矩阵的定义如下

"对于一个 n 阶实方阵 A, 若方阵 B 使得 $AB = BA = I_n$, 则称 B 是 A 的逆矩阵."

事实上, 该定义中的两个条件有一个是多余的。试证明如下事实:

"若 A 是一个 $n\times n$ 实方阵, 若方阵 B 使得 $AB = I_n$, 则必有 $BA = I_n$ "。

证明

$$
\text{由于 } \operatorname{rank}(AB) \leqslant \operatorname{rank}(B)
$$

$$
\operatorname{rank}(B) = n
$$

$$
\forall x, \quad \exists y \quad \text{使} x = By
$$

$$
BAx = BABy = By = x
$$

$$
\text{所以} BA = I
$$

✍ **练习 4.9** 本练习题利用特征根证明矩阵不等式。首先介绍一些记号或定义:

(1) $A \geq 0$ (A 是半正定矩阵); $A > 0$ (A 是正定矩阵); $A \geq B \Leftrightarrow A - B \geq 0$; $A > B \Leftrightarrow A - B > 0$ 。

(2) 定义平方根矩阵: 若 $A_{n\times n} \geq 0$ 的谱分解为 $A = O\Lambda O^\top$, 其中对角阵 $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ 的对角元为 A 的特征根, O 是正交矩阵, 则半正定矩阵的 (唯一) 平方根定义为 $A^{1/2} = O\Delta O^\top$, 其中 $\Delta = \operatorname{diag}\left(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}\right)$ 。定义 $A^{-1/2} = \left(A^{1/2}\right)^{-1}$ 。证明如下事实:

(a) 以 $\lambda(A)$ 表示 A 的任一特征根, 则 $A \leq I_n \Leftrightarrow \lambda(A) \leq 1$, 且 $A \geq I_n \Leftrightarrow \lambda(A) \geq 1$.

(b) 若 $0 < A \leq I_n$, 则 $A^{-1} \geq I_n$.

(c) 若 $A \geq B > 0$, 则对任何 $k \times n$ 矩阵 C 有 $CAC^\top \geq CAC^\top \geq 0$, 特别地, $B^{-1/2}AB^{-1/2} \geq I_n$.

(d) 若 $A \geq B > 0$, 则 $B^{-1} \geq A^{-1} > 0$ 。

(e) (挑战) 若 $A \geq B > 0$, 则 $A^{1/2} \geq B^{1/2} > 0$ 。举例说明 $A \geq B \geq 0$ 未必蕴含 $A^2 \geq B^2$ 。

证明

(a) 由谱分解 $A = O\Lambda O^\top, \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$

$$\begin{aligned}
&A \leqslant I \\
\Leftrightarrow &I - A \geqslant 0 \\
\Leftrightarrow &O \begin{pmatrix} 1-\lambda_1 & & \\ & \ddots & \\ & & 1-\lambda_n \end{pmatrix} O^\top \geqslant 0 \\
\Leftrightarrow &1 - \lambda_i \geqslant 0 \quad i = 1, \ldots n \\
\Leftrightarrow &\lambda(A) \leqslant 1
\end{aligned}$$

同理 $A \geqslant I \Leftrightarrow \lambda(A) \geqslant 1$.

(b) $\quad 0 < A \leqslant I$

$$I - A = O \begin{pmatrix} 1-\lambda_1 & & \\ & \ddots & \\ & & 1-\lambda_n \end{pmatrix} O^\top \geqslant 0$$

由于 $\lambda_i \geqslant 1$ 和 $\lambda_i > 0$ 则 $\frac{1}{\lambda_i} \geqslant 1$

$$A^{-1} - I = O \begin{pmatrix} \frac{1}{\lambda_1} - 1 & & \\ & \ddots & \\ & & \frac{1}{\lambda_n} - 1 \end{pmatrix} O^\top \geqslant 0$$

(c) $\quad A \geqslant B > 0$

$$A - B = O \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} O^\top \geqslant 0$$

$$\begin{aligned}
CAC^\top - CBC^\top &= C(A-B)C^\top \\
&= CO \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} O^\top C^\top \geqslant 0
\end{aligned}$$

$$\text{令} B = I, C = B^{-\frac{1}{2}}$$

$$\text{则} B^{-\frac{1}{2}} A B^{-\frac{1}{2}} \geqslant I$$

(d) $\quad$ 由 (c)

$$\begin{aligned}
&B^{-\frac{1}{2}} A B^{-\frac{1}{2}} \geqslant I \\
&\lambda\left(B^{-\frac{1}{2}} A B^{-\frac{1}{2}}\right) \geqslant 1 \\
\Rightarrow &0 < \lambda\left(B^{\frac{1}{2}} A^{-1} B^{\frac{1}{2}}\right) \leqslant 1 \\
\Rightarrow &0 < B^{\frac{1}{2}} A^{-1} B^{\frac{1}{2}} \leqslant I
\end{aligned}$$

再由 (c)

$$CB^{\frac{1}{2}} A^{-1} B^{\frac{1}{2}} C^\top \leqslant CC^\top$$

取 $C = B^{-\frac{1}{2}}$ 则 $A^{-1} \leqslant B^{-1}$

(e)

$A \geqslant B$, 由 (c) $\quad B^{-\frac{1}{2}}AB^{-\frac{1}{2}} \geqslant I$

$$\forall x, \text{满足} \|x\| = 1$$

$$x^\top B^{-\frac{1}{2}}AB^{-\frac{1}{2}}x = \left\|A^{\frac{1}{2}}B^{-\frac{1}{2}}x\right\|^2 \geqslant \|x\|^2 = 1$$

$$\Rightarrow \left\|A^{\frac{1}{2}}B^{-\frac{1}{2}}x\right\| \geqslant 1$$

$$\Rightarrow A^{\frac{1}{2}}B^{-\frac{1}{2}} \text{的特征值大于 } 1$$

$$\text{由于} \left(A^{\frac{1}{2}}B^{-\frac{1}{4}}\right)B^{-\frac{1}{4}} \text{与} B^{-\frac{1}{4}}\left(A^{\frac{1}{2}}B^{-\frac{1}{4}}\right) \text{有相同特征值}$$

$$\Rightarrow B^{-\frac{1}{4}}A^{\frac{1}{2}}B^{-\frac{1}{4}} \text{的特征值大于 } 1$$

$$\Rightarrow \quad B^{-\frac{1}{4}}A^{\frac{1}{2}}B^{-\frac{1}{4}} \geqslant I$$

$$\text{由 (c)} \Rightarrow A^{\frac{1}{2}} \geqslant B^{\frac{1}{4}}B^{\frac{1}{4}} = B^{\frac{1}{2}}$$

反例:

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \quad B = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

$$A - B = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \geq 0$$

$$A^2 - B^2 = \begin{pmatrix} 3 & 1 \\ 1 & 0 \end{pmatrix}$$

$$det(A^2 - B^2) = -1$$

4.9 HW 9

作业 9 链接

练习 4.1 假设 $y_1, \ldots, y_n \text{iid} \sim \left(\mu, \sigma^2\right), \mu, \sigma^2$ 是未知参数。记样本均值和样本方差分别为 $\bar{y}$ 和 s^2 。令 $\mathbf{y} = (y_1, \ldots, y_n)^\top, \boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top, \mathbb{1} = (1, \ldots, 1)^\top$, 则上述模型可表示为矩阵向量形式

$$\mathbf{y} = \mathbb{1}\mu + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \left(0, \sigma^2 I_n\right).$$

求 $\mathbf{y}$ 在 $C(\mathbb{1})$ 上的投影 $\hat{\mathbf{y}}$ 及 μ 的 LS 估计, 以及 σ^2 的 LS 估计. (提示: $P_{\mathbb{1}}\mathbf{y} = \mathbb{1}\hat{\mu}$)

解 利用投影矩阵:

$$\hat{y} = P_{\mathbb{1}}(y) = \mathbb{1}\left(\mathbb{1}^{\mathbf{T}}\mathbb{1}\right)^{-}\mathbb{1}^{\mathbf{T}}y = \mathbb{1}\bar{y}.$$

已知 $P_{\mathbb{1}}\boldsymbol{y} = \mathbb{1}\hat{\mu}$, 因此 μ 的 LS 估计为 $\bar{y}$ 。方差的 LS 估计为:

$$\hat{\sigma^2} = \frac{RSS}{n-p} = \frac{\sum_{i=1}^n (y-\bar{y})^2}{n-1} = s^2.$$

练习 4.2 假设随机样本数据 $(x_i, y_i), i = 1, \ldots, n$ 满足简单线性回归模型

$$y_i = a + bx_i + \epsilon_i, \epsilon_i, i = 1, \ldots, \text{ iid } \sim N\left(0, \sigma^2\right), \text{且} x_i \text{ 与} \epsilon_i \text{ 独立}, i = 1, \ldots, n.$$

将模型写成矩阵-向量的形式

$$\mathbf{y} = X\beta + \boldsymbol{\epsilon} = \mathbb{1}a + \mathbf{x}b + \boldsymbol{\epsilon}$$

其中设计阵 $X = (\mathbb{1}, \mathbf{x}), \mathbf{x} = (x_1, \ldots, x_n)^\top, \beta = (a, b)^\top$.

(a) 记 $\bar{x} = (x_1 + \ldots + x_n)/n = \mathbf{x}^\top\mathbb{1}/n$. 证明 $C(X)$ 的投影矩阵

$$P_X = \frac{\mathbb{1}\mathbb{1}^\top}{n} + \frac{(\mathbf{x} - \mathbb{1}\bar{x})(\mathbf{x} - \mathbb{1}\bar{x})^\top}{\|\mathbf{x} - \mathbb{1}\bar{x}\|^2}.$$

(提示: 先将 $\mathbf{x}$ 和 $\mathbb{1}$ 正交化).

(b) 求 $\mathbf{y}$ 在 $C(X)$ 上的投影 $\hat{\mathbf{y}} = P_X\mathbf{y}$ 。投影表达式决定了参数 a, b 的 LS 估计: 若 $\hat{\mathbf{y}} = \mathbb{1}\xi + \mathbf{x}\eta$, 则 $\mathbb{1}, \mathbf{x}$ 的系数即是 LS 估计 $\hat{a} = \xi, \hat{b} = \eta$, 试由 $\hat{\mathbf{y}}$ 的表达式求出 $\hat{a}, \hat{b}$.

解

1. 先将 x 正交化:

$$\boldsymbol{x}^\perp = x - P_{\mathbb{1}}x = x - \mathbb{1}\bar{x}$$

已知 $P_x = P_{\mathbb{1}} + P_{x^\perp}$, 因此:

$$P_x = P_{\mathbb{1}} + P_{x^\perp} = \frac{\mathbf{1}\mathbf{1}^{\boldsymbol{T}}}{n} + \frac{(\boldsymbol{x} - \mathbb{1}\bar{x})(\boldsymbol{x} - \mathbb{1}\bar{x})^T}{\|\boldsymbol{x} - \mathbb{1}\bar{x}\|^2}$$

2.

$$\begin{aligned}\hat{y} = P_x(y) &= \left(\frac{\mathbf{1}\mathbf{1}^{\boldsymbol{T}}}{n} + \frac{(\boldsymbol{x} - \mathbb{1}\bar{x})(\boldsymbol{x} - \mathbb{1}\bar{x})^T}{\|\boldsymbol{x} - \mathbb{1}\bar{x}\|^2}\right)y \\ &= \mathbb{1}\left(\frac{\mathbb{1}^{\boldsymbol{T}}y}{n} - \bar{x}\frac{(\boldsymbol{x} - \mathbb{1}\bar{x})^T y}{\|\boldsymbol{x} - \mathbb{1}\bar{x}\|^2}\right) + \boldsymbol{x}\frac{(\boldsymbol{x} - \mathbb{1}\bar{x})^T y}{\|\boldsymbol{x} - \mathbb{1}\bar{x}\|^2}\end{aligned}$$

若 $\hat{\boldsymbol{y}} = \mathbb{1}\xi + \boldsymbol{x}\eta$, 则:

$$\hat{a}=\xi=\frac{\mathbb{1}^{\boldsymbol{T}}y}{n}-\bar{x}\frac{(\boldsymbol{x}-\mathbb{1}\bar{x})^{T}y}{\|\boldsymbol{x}-\mathbb{1}\bar{x}\|^{2}}=\bar{y}-\bar{x}\eta$$
$$\hat{b}=\eta=\frac{(\boldsymbol{x}-\mathbb{1}\bar{x})^{T}y}{\|\boldsymbol{x}-\mathbb{1}\bar{x}\|^{2}}=\frac{S_{yy}}{S_{xx}}$$

练习 4.3 假设线性模型 $y_i=\beta_0+\mathbf{x}_i^\top\mathbf{b}+\epsilon_i, i=1,2,\ldots,n$, 其矩阵-向量形式为

$$\mathbf{y}_{n\times 1}=X_{n\times p}\beta_{p\times 1}+\boldsymbol{\epsilon}=\mathbb{1}\beta_0+Z\boldsymbol{b}+\boldsymbol{\epsilon},$$

其中 $\beta=\left(\beta_0,\mathbf{b}^\top\right)^\top, X=(\mathbb{1},Z), Z=(\mathbf{x}_1,\ldots,\mathbf{x}_n)^\top$ 为所有自变量按行排成的 $n\times(p-1)$ 矩阵。

(a) 假设 $Z^\top\mathbb{1}=0$ (即 Z 的各个列向量都与 $\mathbb{1}$ 正交, 每列之和为 0), 试利用 LS 估计公式 $\hat{\beta}=\left(X^\top X\right)^{-}X^\top\mathbf{y}$ 证明 $\mathbf{b}$ 的 LS 估计

$$\hat{\mathbf{b}}=\left(Z^\top Z\right)^{-}Z^\top\mathbf{y}$$

(b) 试利用投影 $P_X\mathbf{y}=P_{\mathbb{1}}\mathbf{y}+P_Z\mathbf{y}=\mathbb{1}\hat{\beta}_0+Z\hat{\mathbf{b}}$, 说明 $\hat{\mathbf{b}}=\left(Z^\top Z\right)^{-}Z^\top\mathbf{y}$.

(c) 当 $Z^\top\mathbb{1}\neq 0$ 时, 记 Z 的中心化矩阵 $Z_c=Z-P_{\mathbb{1}}Z=Z-\mathbb{1}^\top\overline{\mathbf{x}}^\top$, 其中 $\overline{\mathbf{x}}=Z^\top\mathbb{1}/n=(\mathbf{x}_1+\ \ldots+\mathbf{x}_n)/n$ 为自变量的样本均值。试利用 LS 估计的表达 $\hat{\beta}=\left(\hat{\beta}_0,\hat{\mathbf{b}}^\top\right)^\top=\left(X^\top X\right)^{-1}X^\top\mathbf{y}$ 和分块矩阵求逆公式（第 4 讲命题 1, P12）证明

$$\hat{\mathbf{b}}=\left(Z_c^\top Z_c\right)^{-1}Z_c^\top\mathbf{y}$$

(d) 将 $Z=Z_c+\mathbb{1}\overline{\mathbf{x}}^\top$ 代入模型 (1)

$$\mathbf{y}=\mathbb{1}\beta_0+\left(Z_c+\mathbb{1}\overline{\mathbf{x}}^\top\right)\mathbf{b}+\boldsymbol{\epsilon}=\mathbb{1}\beta_0^*+Z_c\mathbf{b}+\boldsymbol{\epsilon}$$

其中 $\beta_0^*=\beta_0+\overline{\mathbf{x}}^\top\mathbf{b}$, 注意 $Z_c^\top\mathbb{1}=0$, 因此投影 $\hat{\mathbf{y}}=P_X\mathbf{y}=P_{\mathbb{1},Z_c}\mathbf{y}=P_{\mathbb{1}}\mathbf{y}+P_{Z_c}\mathbf{y}$, 根据该表达说明 $\mathbf{b}$ 的最小二乘估计为 $\hat{\mathbf{b}}=\left(Z_c^\top Z_c\right)^{-1}Z_c^\top\mathbf{y}$.

(e) 记

$$S_{\mathbf{xx}}=\frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{x}_i-\overline{\mathbf{x}}\right)\left(\mathbf{x}_i-\overline{\mathbf{x}}\right)^\top, S_{\mathbf{x}y}=\frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{x}_i-\overline{\mathbf{x}}\right)y_i$$

分别为样本方差和样本协方差矩阵, 验证 (c) 或 (d) 中的 LS 估计 $\hat{\mathbf{b}}=S_{\mathbf{xx}}^{-1}S_{\mathbf{x}y}$.

证明

1. 将 $\hat{\beta}=\left(X^TX\right)^{-}X^Ty$ 展开,

$$\begin{pmatrix}\hat{\beta}_0\\ \hat{b}\end{pmatrix}=\begin{pmatrix}\mathbb{1}^{\boldsymbol{T}}\mathbb{1} & \mathbb{1}^{\boldsymbol{T}}Z\\ Z^T\mathbb{1} & Z^TZ\end{pmatrix}^{-}\begin{pmatrix}\mathbb{1}^{\boldsymbol{T}}\\ Z^T\end{pmatrix}y$$

且我们知道 $Z^T\mathbb{1}=0$, 上式化简为:

$$\begin{pmatrix}\hat{\beta}_0\\ \hat{b}\end{pmatrix}=\begin{pmatrix}\mathbb{1}^{\boldsymbol{T}}\mathbb{1} & 0\\ 0 & Z^TZ\end{pmatrix}^{-}\begin{pmatrix}\mathbb{1}^{\boldsymbol{T}}\\ Z^T\end{pmatrix}y=\begin{pmatrix}\frac{\mathbb{1}^{\boldsymbol{T}}y}{n}\\ \left(Z^TZ\right)^{-}Z^Ty\end{pmatrix}$$

即得 $\hat{b}$ 的 LS 估计。

2. 已知 $Z^T\mathbb{1}=0$, 所以

$$P_x=P_{\mathbb{1}}+P_Z=\frac{\mathbf{1}\mathbf{1}^{\boldsymbol{T}}}{n}+Z\left(Z^TZ\right)^{-}Z^T$$

因此:

$$P_X \boldsymbol{y} = P_{\mathbb{1}} + P_Z(y) = \mathbb{1}\bar{y} + Z\left(Z^T Z\right)^{-} Z^T y = \mathbb{1}\hat{\beta}_0 + Z\hat{b}$$

于是 $Z\left(\left(Z^T Z\right)^{-} Z^T y - \hat{b}\right) = 0$, 由于样本彼此应不同, 矩阵列满秩, 故: $\hat{b} = \left(Z^T Z\right)^{-} Z^T y$.

3.

$$\begin{aligned}
\begin{pmatrix} \hat{\beta}_0 \\ \hat{b} \end{pmatrix} &= \begin{pmatrix} \mathbb{1}^{\boldsymbol{T}}\mathbb{1} & \mathbb{1}^{\boldsymbol{T}} Z \\ Z^T\mathbb{1} & Z^T Z \end{pmatrix}^{-1} \begin{pmatrix} \mathbb{1}^{\boldsymbol{T}} \\ Z^T \end{pmatrix} y \\
&= \begin{pmatrix} * & * \\ -\left(Z^T Z - n\bar{x}\bar{x}^T\right)^{-1}\bar{x} & \left(Z^T Z - n\bar{x}\bar{x}^T\right)^{-1} \end{pmatrix} \begin{pmatrix} \mathbb{1}^{\boldsymbol{T}} \\ Z^T \end{pmatrix} y \\
&= \begin{pmatrix} * \\ \left(Z^T Z - n\bar{x}\bar{x}^T\right)^{-1}\left(Z^T - \bar{x}\mathbb{1}^T\right) y \end{pmatrix}
\end{aligned}$$

同时有:

$$\left(Z_c^T Z_c\right)^{-1} Z_c^T = \left(Z^T Z - n\bar{x}\bar{x}^T\right)^{-1}\left(Z^T - \bar{x}\mathbb{1}^T\right)$$

因此: $\hat{b} = \left(Z_c^T Z_c\right)^{-1} Z_c^T y$.

4. 利用中心化, 我们得到 $Z_c^T \mathbb{1} = 0$. 与第 2 问相同, 即得:

$$\hat{b} = \left(Z_c^T Z_c\right)^{-1} Z_c^T y$$

5. 事实上, $Z_c^T Z_c = \sum (x_i - \bar{x})(x_i - \bar{x})^T = (n-1)S_{xx}$, $Z_c^T y = (n-1)S_{xy}$. 故: $\hat{b} = \left(Z_c^T Z_c\right)^{-1} Z_c^T y = S_{xx}^{-1} S_{xy}$.

注 分块矩阵求逆公式，参见 Week04 第 12 页。

✍ **练习 4.4** 下面叙述的是简单回归模型中求解工具变量最小二乘估计的方法步骤:

假设随机变量 x, y 满足总体模型 $y = a + bx + \epsilon, \epsilon \sim \left(0, \sigma^2\right)$, 但其中 x 与 ϵ 不独立 (内生)。假设存在一个随机变量 z, 它与 ϵ 独立 (外生), 特别地, $\operatorname{cov}(\epsilon, z) = 0$, 即

$$0 = \operatorname{cov}(\epsilon, z) = \operatorname{cov}(y - a - bx, z) = \Sigma_{yz} - b\Sigma_{xz} \Rightarrow b = \Sigma_{yz}/\Sigma_{xz}$$

假设样本数据为 $(y_i, x_i, z_i), i = 1, \ldots, n$, 为了估计 b 我们应用矩估计方法, 在上述表达中代入样本协方差

$$S_{yz} = \sum_{i=1}^{n} (z_i - \bar{z})(y_i - \bar{y})/(n-1), S_{xz} = \sum_{i=1}^{n} (z_i - \bar{z})(x_i - \bar{x})/(n-1)$$

即得 b 的估计

$$\tilde{b} = S_{yz}/S_{xz} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}$$

该估计称为工具变量最小二乘估计。可以证明该估计是 b 的渐近无偏估计 (此处略去)。

从 b 或 $\tilde{b}$ 的表达式来看，上述讨论中对于 z 遗漏了什么条件?

解 还需要 z 与 x 相关。因为当 $\Sigma_{xz} = 0$ 时，b 无解。

4.10 HW 10

作业 10 链接

练习 4.1 对于一般线性模型 $\mathbf{y} = X\beta + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \left(\mathbf{0}, \sigma^2 I_n\right)$, 我们知道 $\left(I_n - P_X\right) X = 0$, 其中 $P_X = X\left(X^\top X\right)^- X^\top$ 是投影阵, 所以

$$RSS = \mathbf{y}^\top \left(I_n - P_X\right) \mathbf{y} = \boldsymbol{\epsilon}^\top \left(I_n - P_X\right) \boldsymbol{\epsilon},$$

即 RSS 不依赖于 $\boldsymbol{\beta}$ 。基于上述观点重新证明第 6 讲引理 1(2) : 对于简单线性模型, 残差平方和 $RSS = s_{\epsilon\epsilon} - s_{\epsilon x}^2 / s_{xx}$ 。

解

简单线性模型

$$y_i = a + bx_i + \varepsilon_i, i = 1, 2, \cdots, n$$

其矩阵-向量形式为 $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, 其中

$$X = \begin{pmatrix} \mathbf{1} & \boldsymbol{x} \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} a \\ b \end{pmatrix}$$

设

$$\tilde{X} = \begin{pmatrix} \mathbf{1} & \frac{1}{\sqrt{s_{xx}}}(\boldsymbol{x} - \mathbf{1}\bar{x}) \end{pmatrix}$$

则 $C(X) = C(\tilde{X})$

$$P_X = P_{\tilde{X}} = \tilde{X}(\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T = \left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{s_{xx}}\right)_{n\times n}$$

所以

$$\begin{aligned}
RSS &= \varepsilon^T \left(I_n - P_X\right) \varepsilon \\
&= \varepsilon^T \varepsilon - \varepsilon^T P_X \varepsilon \\
&= \sum_{i=1}^n \varepsilon_i^2 - \sum_{i,j=1}^n \varepsilon_i \varepsilon_j \left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{s_{xx}}\right) \\
&= \sum_{i=1}^n \varepsilon_i^2 - \sum_{i,j=1}^n \frac{\varepsilon_i \varepsilon_j}{n} - \frac{1}{s_{xx}} \sum_{i,j=1}^n \varepsilon_i \varepsilon_j (x_i - \bar{x})(x_j - \bar{x}) \\
&= \sum_{i=1}^n \varepsilon_i^2 - \frac{1}{n}\left(\sum_{i=1}^n \varepsilon_i\right)^2 - \frac{1}{s_{xx}}\left(\sum_{i=1}^n \varepsilon_i (x_i - \bar{x})\right)^2 \\
&= s_{\varepsilon\varepsilon} - \frac{s_{\varepsilon x}^2}{s_{xx}}
\end{aligned}$$

注 注意 $s_{\varepsilon\varepsilon}$ 是中心化的形式!

练习 4.2 假设数据 $(y_i, x_i, z_i), i = 1, \ldots, n$ 满足线性模型

$$y_i = a + bx_i + cz_i + \epsilon_i, \epsilon_i \sim \left(0, \sigma^2\right), \epsilon_i \text{ 与} x_i, z_i \text{ 独立}$$

或等价地 $\left(\mathbf{x} = (x_1, \ldots, x_n)^\top, \mathbf{z} = (z_1, \ldots, z_n)^\top\right)$

$$\mathbf{y} = \mathbb{1}a + \mathbf{x}b + \mathbf{z}c + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \left(0, \sigma^2 I_n\right), \quad \boldsymbol{\epsilon} \perp (\mathbf{x}, \mathbf{z}).$$

(a) 记 $\hat{c}_0 = S_{xz}/S_{zz}, S_{xz} = \frac{1}{n-1}\sum (x_i - \bar{x})(z_i - \bar{z}), S_{zz} = \frac{1}{n-1}\sum (z_i - \bar{z})^2$, 试用第 11 讲命题 4 证明 1 的投影方法 (即先将三个向量 $\mathbb{1}, \mathbf{x}, \mathbf{z}$ 正交化, 再投影) 而不是直接应用命题 4 的结论, 证明 b 的 LS 估计

$$\hat{b} = \frac{\sum \left[(x_i - \bar{x} - (z_i - \bar{z})\hat{c}_0\right] y_i}{\sum \left[(x_i - \bar{x} - (z_i - \bar{z})\hat{c}_0\right]^2}.$$

(注: 该表达式表明, 为了求解 b 的 LS 估计, 我们可以分两步回归: (1) $\mathrm{lm}(x \sim z) : x_i = a_0 + c_0 z_i + \delta_i$, 求得回归系

数 $\hat{c}_0 = s_{xz}/s_{zz}$ 以及 $\hat{a}_0 = \bar{x} - \bar{z}\hat{c}_0$, 得到残差 $x_i^{\perp} = x_i - \bar{x} - (z_i - \bar{z})\,\hat{c}_0$. (2) $y \sim x^{\perp} : y_i = \alpha + \beta x_i^{\perp} +$ error, 由此得到的斜率 LS 估计 $\hat{\beta}$ 就是 $\hat{b}$)

(b) 证明

$$R^2 = \frac{r_{yx}^2 + r_{yz}^2 - 2r_{yx}r_{yz}r_{xz}}{1 - r_{xz}^2},$$

其中 r_{xz} 为两个自变量之间的样本相关系数。

证明 记

$$s_{xy} = \sum (x_i - \bar{x})\,(y_i - \bar{y})$$

对 $X = \begin{pmatrix} \mathbf{1} & \boldsymbol{x} & \boldsymbol{z} \end{pmatrix}$ 正交化得到:

于是

$$\hat{\boldsymbol{y}} = P_X \boldsymbol{y} = P_{\tilde{X}} \boldsymbol{y} = \mathbf{1}\hat{a} + \boldsymbol{x}\hat{b} + \boldsymbol{z}\hat{c}$$

关注 x 的系数, 可得

$$\begin{aligned}
\hat{b} &= \left[\frac{\boldsymbol{x}^T - \bar{x}\mathbf{1}^T}{s_{xx}} - \frac{\frac{s_{xz}}{s_{xx}}\left[\boldsymbol{z}^T - \mathbf{1}\bar{z} - \frac{s_{xz}}{s_{xx}}(\boldsymbol{x} - \mathbf{1}\bar{x})^T\right]}{s_{zz} - s_{zx}^2/s_{xx}}\right]\boldsymbol{y} \\
&= \frac{s_{xy}}{s_{xx}} - \frac{s_{xz}\left(s_{zy} - \frac{s_{xz}}{s_{xx}}s_{xy}\right)}{s_{zz}s_{xx} - s_{xz}^2} \\
&= \frac{s_{xy}s_{xx}s_{zz} - s_{xy}s_{xz}^2 - s_{xz}s_{zy}s_{xx} + s_{xy}s_{xz}^2}{s_{xx}\left(s_{zz}s_{xx} - s_{xz}^2\right)} \\
&= \frac{s_{xy}s_{zz} - s_{xz}s_{zy}}{s_{zz}s_{xx} - s_{xz}^2} = \frac{s_{xy} - s_{zy}}{s_{xx} - s_{xz}\hat{c}_0} = \frac{\sum\left[x_i - \bar{x} - (z_i - \bar{z})\,\hat{c}_0\right]y_i}{\sum\left[x_i - \bar{x} - (z_i - \bar{z})\,\hat{c}_0\right]^2}
\end{aligned}$$

同时, x 和 z 地位对等，交换位置即有

$$\hat{c} = \frac{s_{zy}s_{xx} - s_{xz}s_{xy}}{s_{zz}s_{xx} - s_{xz}^2}$$

考虑 $R^2 = \operatorname{var}(\hat{\boldsymbol{y}})/\operatorname{var}(\boldsymbol{y}) = (n-1)\operatorname{var}(\hat{\boldsymbol{y}})/s_{yy}$,

$$\begin{aligned}
\operatorname{var}(\hat{\boldsymbol{y}}) &= \operatorname{var}(\boldsymbol{x}\hat{b} + \boldsymbol{z}\hat{c}) \\
&= \operatorname{var}(\boldsymbol{x})\hat{b}^2 + \operatorname{var}(\boldsymbol{z})\hat{c}^2 + 2\hat{b}\hat{c}\operatorname{cov}(\boldsymbol{x}, \boldsymbol{z}) \\
&= \frac{1}{n-1}\left(s_{xx}\hat{b}^2 + s_{zz}\hat{c}^2 + 2s_{xz}\hat{b}\hat{c}\right) \\
&= \frac{1}{n-1}\frac{s_{xx}\left(s_{xy}s_{zz} - s_{xz}s_{zy}\right)^2 + s_{zz}\left(s_{zy}s_{xx} - s_{xz}s_{xy}\right)^2 + 2s_{xz}\left(s_{zy}s_{xx} - s_{xz}s_{xy}\right)\left(s_{xy}s_{zz} - s_{xz}s_{zy}\right)}{\left(s_{zz}s_{xx} - s_{xz}^2\right)^2} \\
&= \frac{1}{n-1}\frac{s_{xy}^2 s_{zz} + s_{yz}^2 s_{xx} - 2s_{xy}s_{yz}s_{xz}}{s_{zz}s_{xx} - s_{xz}^2} \\
R^2 &= \frac{s_{xy}^2 s_{zz} + s_{yz}^2 s_{xx} - 2s_{xy}s_{yz}s_{xz}}{\left(s_{zz}s_{xx} - s_{xz}^2\right)s_{yy}}
\end{aligned}$$

注意到

$$r_{xy}^2 = \frac{s_{xy}^2}{s_{xx}s_{yy}}$$

于是

$$R^2 = \frac{r_{yx}^2 + r_{yz}^2 - 2r_{yx}r_{yz}r_{xz}}{1 - r_{xz}^2}$$

练习 4.3 假设线性模型 $\mathbf{y} = X_1\beta_1 + X_2\beta_2 + \boldsymbol{\epsilon}$, 其中 X_1 为 $n \times (p-q)$ 矩阵, X_2 为 $n \times q$ 矩阵, β_2 为 $q \times 1$ 参数向量。记 $X_2^{\perp} = X_2 - P_{X_1}X_2$ 。假设 W 是任何 $n \times q$ 矩阵使得 $W^{\top}X_1 = 0, W^{\top}X_2$ 可逆, 证明 $\tilde{\beta}_2 = \left(W^{\top}X_2\right)^{-1}W^{\top}\mathbf{y}$ 是 β_2 的无偏估计, 并证明 $\operatorname{var}\left(\tilde{\beta}_2 \mid X\right) \geq \sigma^2\left(X_2^{\perp}X_2^{\perp}\right)^{-1}$, 当 $W = X_2^{\perp}$ 时等号成立。

证明

$$\begin{aligned} E\left(\tilde{\beta}_2 \mid X\right) &= \left(W^T X_2\right)^{-1} W^T \left(X_1\beta_1 + X_2\beta_2\right) \\ &= \left(W^T X_2\right)^{-1} W^T X_2 \beta_2 = \beta_2 \end{aligned}$$

因此: $E\left(\tilde{\beta}_2\right) = E\left(E\left(\tilde{\beta}_2 \mid X\right)\right) = \beta_2$.

记 $C = \left(W^T X_2\right)^{-1} W^T$, 则 $CX_2 = CX_2^{\perp} = I$, 根据投影阵 $P \leqslant I$, $W^T X_2 = W^T X_2^{\perp}$

$$\begin{aligned} \operatorname{var}\left(\tilde{\beta}_2 \mid X\right) &= \sigma^2 CC^T \\ \left({X_2^{\perp}}^T X_2^{\perp}\right)^{-1} &= CX_2^{\perp}\left({X_2^{\perp}}^T X_2^{\perp}\right)^{-1} {X_2^{\perp}}^T C^T \\ &= CP_{X_2^{\perp}} C^T \leqslant CC^T \\ \Rightarrow \operatorname{var}\left(\tilde{\beta}_2 \mid X\right) &\geqslant \sigma^2 \left({X_2^{\perp}}^T X_2^{\perp}\right)^{-1} \end{aligned}$$

练习 4.4 假设 A 为任一 $n \times p$ 列满秩矩阵, 其各列为 $A = (\mathbf{a}_1, \ldots, \mathbf{a}_p)$, 我们知道 $A^\top A$ 的 (i,j) 元为 $\mathbf{a}_i^\top \mathbf{a}_j$ 。记 $\mathbf{a}_i^{\perp} = \mathbf{a}_i - P_{A_{(-i)}}\mathbf{a}_i$, 其中 $A_{(-i)}$ 为 A 删除第 i 列后剩余的列组成的矩阵。证明 $\left(A^\top A\right)^{-1}$ 的 (i,j) 元等于 $\frac{\mathbf{a}_i^{\perp\top}\mathbf{a}_j^{\perp}}{\|\mathbf{a}_i^{\perp}\|^2\|\mathbf{a}_j^{\perp}\|^2}$. 提示: 假设线性模型 $\mathbf{y} = A\beta + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \left(0, \sigma^2 I_n\right), \beta = (\beta_1, \ldots, \beta_p)^\top$, 注意这里 A 的第一列未必是 $\mathbb{1}$, 但 LS 估计仍具有形式 $\hat{\beta} = \left(A^\top A\right)^{-1} A^\top \mathbf{y}$, 方差矩阵为 $\operatorname{var}(\hat{\beta} \mid A) = \sigma^2\left(A^\top A\right)^{-1}$, 其 (i,j) 元为 $\hat{\beta}_i, \hat{\beta}_j$ 的协方差, 因此只需计算 $\operatorname{cov}\left(\hat{\beta}_i, \hat{\beta}_j \mid A\right)$ (利用 11 讲命题 4 给出 $\hat{\beta}_i, \hat{\beta}_j$ 的表达).

解

线性模型可以写为: $\boldsymbol{y} = a_1\beta_1 + \cdots + a_p\beta_p + \epsilon$. 由前已知 β_i 的 LS 估计为 $\hat{\beta}_i = \left(a_i^{\perp T} a_i^{\perp}\right)^{-1} a_i^{\perp T} y$, 于是:

$$\begin{aligned} \operatorname{Cov}\left(\hat{\beta}_i, \hat{\beta}_j \mid A\right) &= \left({a_i^{\perp}}^T a_i^{\perp}\right)^{-1} {a_i^{\perp}}^T \operatorname{Var}(y \mid A) a_j^{\perp}\left({a_j^{\perp}}^T a_j^{\perp}\right)^{-1} \\ &= \sigma^2 \frac{{a_i^{\perp}}^T a_j^{\perp}}{\left\|a_i^{\perp}\right\|^2 \left\|a_j^{\perp}\right\|^2} \end{aligned}$$

所以 $\left(A^T A\right)^{-1}$ 的 (i,j) 元为 $\frac{{a_i^{\perp}}^T a_j^{\perp}}{\|a_i^{\perp}\|^2\|a_j^{\perp}\|^2}$.

4.11 HW 11

作业 11 链接

本次作业，很多同学用的方法相当复杂，没有选出简单的公式。涉及 F 检验的内容，对于具体的模型，核心是把全模型和零模型的简单形式写出来，计算 $\hat{y}$ 与 $\hat{y}_0$, 再考虑问题，具体做法参见解答。一味地套原始公式会面对复杂的矩阵，难以处理！

练习 4.1 假设 n 个个体被随机分配服用某种药物, 按照剂量从小到达分成 K 个组, 假设服用第 k 种剂量的个体数为 n_k, 响应的均值为 μ_k, 具体如下:

$$y_1, \ldots, y_{n_1} \text{ iid } \sim N\left(\mu_1, \sigma^2\right); \quad y_{n_1+1}, \ldots, y_{n_1+n_2} \text{ iid } \sim N\left(\mu_2, \sigma^2\right) \text{ 等等}$$

以线性模型表示如下:

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim N_n\left(0, \sigma^2 I_n\right), \boldsymbol{\beta} = \left(\mu_1, \mu_2, \ldots, \mu_K\right)^\top,$$

设计阵为

$$X_{n\times K} = \begin{pmatrix} \mathbb{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{0}_{n_2} & \mathbb{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{0}_{n_K} & \mathbf{0}_{n_K} & \cdots & \mathbb{1}_{n_K} \end{pmatrix}$$

其中 $\mathbb{1}_{n_i}$ 代表长度为 n_i 的分量全是 1 的向量, $\mathbf{0}_{n_i}$ 代表长度为 n_i 的分量全是 0 的向量, $n = n_1 + \ldots + n_K, K \geq 2$. 求 $H_0 : \mu_k = k\mu_1 (k = 1, 2, \ldots, K)$ 的 F 检验统计量.

解 由题知, 线性模型表示如下:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \epsilon \sim N_n\left(0, \sigma^2 I_n\right), \beta = \left(\mu_1, \cdots, \mu_K\right)^T$$

设计阵记为 $X_{n\times K} = (g_1, \cdots, g_k)$, 其中 g_k 为列向量。

在 H_0 假设下，模型为 $y = (g_1 + 2g_2 + \cdots + kg_k)\,\mu_1 + \epsilon = Z\mu_1 + \epsilon$ 。因此

$$\hat{\mu_1} = \frac{\sum_{i=1}^m i n_i \overline{y_i}}{\sum_{i=1}^m i^2 n_i}$$

于是

$$\hat{y}_0 = \left(\hat{\mu_1}, \cdots, k\hat{\mu_1}\right)^T$$
$$\hat{y} = P_{g_1} y + \cdots + P_{g_K} y = \left(\overline{y_1}, \cdots, \overline{y_1}, \cdots, \overline{y_k}, \cdots, \overline{y_k}\right)^T$$

将上式代入 F 检验的一般形式: $F = \frac{n-p}{q} \times \frac{\|\hat{y}-\hat{y_0}\|^2}{\|y-\hat{y}\|^2}$ 。即得:

$$F_{K-1,n-K} = \frac{n-K}{K-1} \times \frac{\sum_{i=1}^K \left(\overline{y_i} - i\hat{\mu}_1\right)^2}{\sum_{k=1}^K \sum_{i=1}^{n_k} \left(y_{ki} - \bar{y}_k\right)^2}$$

练习 4.2 为了估计两个物品的重量 α, β, 用天平称量三次, 三次测量分别测的是 $\alpha, \alpha - \beta$ （天平一边放一个物品), $\alpha + \beta$ (两个物品都放天平的同一边), 得到的测量值分别为 y_1, y_2, y_3 。假设天平的测量误差服从 $N\left(0, \sigma^2\right)$ (与被测物品的真实重量无关)。

(a) 写出回归模型, 并求出 α, β, σ^2 的估计。

(b) 求 $H_0 : \alpha = \beta$ 的 F 检验统计量。

解

1. 由题知, 线性模型如下:

$$\boldsymbol{y}=\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}=\begin{pmatrix} 1 & 0 \\ 1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \end{pmatrix}+\boldsymbol{\epsilon}=A\boldsymbol{\eta}+\boldsymbol{\epsilon}$$

由前已知的最小二乘估计，可知:

$$\hat{\boldsymbol{\eta}}=\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix}=\left(A^T A\right)^{-1} A^T y=\begin{pmatrix} \frac{1}{3} & 0 \\ \frac{1}{2} & 0 \end{pmatrix}\begin{pmatrix} 1 & 1 & 1 \\ 0 & -1 & 1 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}=\begin{pmatrix} \frac{1}{3}\left(y_1+y_2+y_3\right) \\ \frac{1}{2}\left(y_3-y_2\right) \end{pmatrix}$$

同理 $\hat{\sigma}^2=\frac{RSS}{n-p}=\frac{\sum_{i=1}^3\|y_i-\hat{y}_i\|^2}{3-2}$ 。其中:

$$y-\hat{y}=\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}-\begin{pmatrix} 1 & 1 & 1 \\ 0 & -1 & 1 \end{pmatrix}\begin{pmatrix} \frac{1}{3}\left(y_1+y_2+y_3\right) \\ \frac{1}{2}\left(y_3-y_2\right) \end{pmatrix}=\begin{pmatrix} \frac{2}{3}y_1-\frac{1}{3}y_2-\frac{1}{3}y_3 \\ -\frac{1}{3}y_1+\frac{1}{6}y_2+\frac{1}{6}y_3 \\ -\frac{1}{3}y_1+\frac{1}{6}y_2+\frac{1}{6}y_3 \end{pmatrix}$$

化简即得: $\hat{\sigma}^2=\frac{1}{6}\left(2y_1-y_2-y_3\right)^2$ 。

2. 求 $H_0:\alpha=\beta$ 的 F 检验统计量: 由前已知, $\hat{y}=(\hat{\alpha},\hat{\alpha}-\hat{\beta},\hat{\alpha}+\hat{\beta})^T$, 即

$$\hat{y}=\begin{pmatrix} \frac{1}{3}\left(y_1+y_2+y_3\right) \\ \frac{1}{3}y_1+\frac{5}{6}y_2-\frac{1}{6}y_3 \\ \frac{1}{3}y_1-\frac{1}{6}y_2+\frac{5}{6}y_3 \end{pmatrix}$$

在零假设下，模型为:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_2 \end{pmatrix}=\begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}\alpha+\epsilon$$

求得 α 在 V_0 下的 LS 估计:

$$\hat{\alpha_0}=\frac{y_1+2y_3}{5}$$

于是:

$$\hat{y}_0=\left(\frac{y_1+2y_3}{5},0,\frac{2y_1+4y_3}{5}\right)^T$$

代入 F 检验的一般形式:

$$F_{1,1}=\frac{n-p}{q}\times\frac{\|\hat{y}-\hat{y_0}\|^2}{\|y-\hat{y}\|^2}=\frac{\|\hat{y}-\hat{y_0}\|^2}{\hat{\sigma^2}}=\frac{\left(2y_1+5y_2-y_3\right)^2}{5\left(2y_1-y_2-y_3\right)^2}$$

其中 $\hat{y}-\hat{y}_0=\left(\frac{1}{15},\frac{1}{6},-\frac{1}{30}\right)^T\{2y_1+5y_2-y_3\}$

练习 4.3 为了比较处理 1 (对照、安慰剂) 和处理 2 (药物) 是否有差异, 我们将研究对象进行配对匹配使得他们尽量相似, 其中一个（随机决定）接受处理 1 , 另一个接受处理 2 , 他们组成一个区组（这里称为配对)。假设有两个配对, 第 j 对研究对象的响应为 (y_{1j},y_{2j}) , $j=1,2$ 其中下标中的 1,2 分别代表处理 1 和 2 。数据列表如下:

假设正态模型: $y_{ij}\sim N\left(\mu_{ij},\sigma^2\right)$, $i,j=1,2$, 其中 $y_{11},y_{21},y_{12},y_{22}$ 独立, 假设处理的效果在两个区组中相同: $\mu_{21}-\mu_{11}=\mu_{22}-\mu_{12}$, 记之为 δ 。

(a) 验证 $H_0:\delta=0$ 的 F 检验 $F=t_{\text{pair}}^2$, 其中 t_{pair} (成对 t 检验统计量) 如下

		区组/配对 1	区组/配对 2
处理	1	y_{11}	y_{12}
	2	y_{21}	y_{22}

$$t_{\text{pair}} = \frac{y_{21}+y_{22}-y_{11}-y_{12}}{|y_{11}+y_{22}-y_{12}-y_{21}|} \overset{H_0}{\sim} t_1$$

(评注: t_1 分布也称作 Cauchy 分布)

(b) 假设 r.v. x 关于 0 对称，$P(x=0)=0, P(\operatorname{sgn}(x)=\pm 1)=1/2$, 其中 $\operatorname{sgn}(x)$ 代表 x 的符号。证明 $x=\operatorname{sgn}(x)|x|$ 中 $\operatorname{sgn}(x)$ 与 $|x|$ 独立。假设 σ 是符号随机变量, $P(\sigma=\pm 1)=1/2$), 且 σ 与 x 独立，证明 $x \overset{d}{=} \sigma|x| \overset{d}{=} \sigma x$, 其中 $\overset{d}{=}$ 表示同分布。

(c) 假设 x_1, x_2 iid $\sim N(0,1)$, 证明 $\frac{x_1}{x_2}$ 与 $\frac{x_1}{|x_2|}$ 分布相同 (都服从 t_1 分布)。这说明 (1) 式中 t_{pair} 统计量也可取为

$$t_{\text{pair}} = \frac{y_{21}+y_{22}-y_{11}-y_{12}}{y_{11}+y_{22}-y_{12}-y_{21}}$$

(d) 假如列表中的数据不是配对数据而是随机样本, 即接受处理 1 的 y_{11}, y_{12} iid $\sim N\left(\mu_1, \sigma^2\right)$, 接受处理 2 的 y_{21}, y_{22} iid $\sim N\left(\mu_2, \sigma^2\right)$, 数据表示如下 (注意中间没有竖线, y_{11} 和 y_{12} 地位是对称的, 即每一行内的数据没有次序):

		id 1	id 2
处理	1	y_{11}	y_{12}
	2	y_{21}	y_{22}

此时 $H_0: \mu_1=\mu_2$ 的检验为两样本 t 检验, 验证

$$t_{two-sample} = \frac{y_{21}+y_{22}-y_{11}-y_{12}}{\sqrt{(y_{11}-y_{12})^2+(y_{21}-y_{22})^2}} \overset{H_0}{\sim} t_2$$

比较 t_{pair} 和 $t_{\text{two-sample}}$ 的差别, 并讨论为什么有这种差别。

证明

1. 成对比较的假设检验: 令 $z_1=y_{21}-y_{11}, z_2=y_{22}-y_{12}$. 因此:

$$t_{\text{pair}} = \frac{\bar{z}}{s_z/\sqrt{n}} = \frac{y_{22}+y_{21}-y_{11}-y_{12}/2}{\sqrt{\sum_{i=1}^2\left(y_{2i}-y_{1i}-(\overline{y_2}-\overline{y_1})\right)^2/2}} = \frac{y_{22}+y_{21}-y_{11}-y_{12}}{|y_{11}+y_{22}-y_{21}-y_{12}|} \sim t_1$$

2. 先证明 $\operatorname{sgn}(x)$ 与 $|x|$ 独立, 已知 x 是对称的随机变量:

$$P(x=a \mid |x|=a) = P(x=-a \mid |x|=a) = \frac{1}{2}, \quad \forall a$$

因此:

$$P(\operatorname{sgn}(x)=1 \mid |x|=t) = P(x=t \mid |x|=t) = \frac{1}{2} = P(\operatorname{Sgn}(x)=1)$$

$\operatorname{Sgn}(x)$ 与 $|x|$ 独立。

可知 $Sgn(x)$ 与 σ 独立同分布, 于是由 $x=Sgn(x)|x|$ 可知, $x \overset{d}{=} \sigma|x|$ 。同样的, $\operatorname{Sgn}(x)\sigma \overset{d}{=} \sigma$, 相互独立，均与 x 独立。所以 $x \overset{d}{=} \sigma|x| = \sigma \operatorname{Sgn}(x) x \overset{d}{=} \sigma x$

3. 我们反复使用上问结论，注意到 x_1, x_2 独立同分布且为对称随机变量:

$$\frac{x_1}{x_2} \stackrel{d}{=} \frac{\operatorname{Sgn}(x_1)|x_1|}{\operatorname{Sgn}(x_2)|x_2|} \stackrel{d}{=} \frac{\sigma|x_1|}{|x_2|} \stackrel{d}{=} \frac{x_1}{|x_2|}$$

4. 由于没有分组，样本地位对称，此为方差相同的两样本 t 检验：
   检验统计量为：

$$T_w = \frac{\bar{Y} - \bar{X}}{S_w}\sqrt{\frac{mn}{m+n}} \sim t_{n+m-2}$$

   即

$$t_{two-sample} = \frac{y_{21} + y_{22} - y_{11} - y_{12}}{\sqrt{2(y_{11} - \overline{y_1})^2 + 2(y_{12} - \overline{y_1})^2 + 2(y_{21} - \overline{y_2})^2 + 2(y_{22} - \overline{y_2})^2}}$$

   化简即得：

$$t_{two-sample} = \frac{y_{21} + y_{22} - y_{11} - y_{12}}{\sqrt{(y_{11} - y_{12})^2 + (y_{21} - y_{22})^2}} \sim t_2$$

   成对比较与两样本 t 检验的区别在于假设的样本情形不同：在后者中，我们假设来自两个正态总体的样本相互独立，但现实情形中，有可能两个正态的样本是来自于同一个总体上的重复观察，于是我们引进了成对比较检验，通过作差消除其他方面的影响，转而研究引起变化的量，譬如药效等。

练习 4.4 (参见第 11 讲例 3) 假设 5 个中心标准化的随机变量 Y, W, U, X, V 的协方差矩阵 (即相关系数矩阵) 如下

		Y Son's occ	W Son's 1st job	U Son's ed	X Dad's occ	V Dad's ed
$\bar{Y}$	Son's occ	1.000	.541	.596	.405	.322
W	Son's 1st job	.541	1.000	.538	.417	.332
U	Son's ed	.596	.538	1.000	.438	.453
X	Dad's occ	.405	.417	.438	1.000	.516
V	Dad's ed	.322	.332	.453	.516	1.000

假设线性模型（因为所有变量已经中心化，故无截距）

$$U = \beta_1 X + \beta_2 V + \epsilon, \epsilon \sim N\left(0, \sigma^2\right)$$

(a) 试求 β_1, β_2, σ 的 LS 估计。
(b) 求该模型的决定系数 R^2，回归方程的显著性检验统计量 F 及其 p 值 ($n = 20000$).
(c) 求 $H_0 : \beta_1 = 0$ 的 t 检验统计量及其 p 值。

注 注意中心化模型的自由度参数！

解

1. 已知线性模型为：$U = \beta_1 X + \beta_2 V + \epsilon, \epsilon \sim N\left(0, \sigma^2\right)$.
   对应的相关系数矩阵为：

$$S = \begin{pmatrix} 1 & 0.438 & 0.453 \\ 0.438 & 1 & 0.516 \\ 0.453 & 0.516 & 1 \end{pmatrix} = \begin{pmatrix} S_{UU} & S_{U\boldsymbol{Z}} \\ S_{\boldsymbol{Z}U} & S_{\boldsymbol{Z}\boldsymbol{Z}} \end{pmatrix}$$

   由前已知的 LS 估计：

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = S_{ZZ}^{-1} S_{ZU} = \begin{pmatrix} 1 & 0.516 \\ 0.516 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0.438 \\ 0.453 \end{pmatrix} = \begin{pmatrix} 0.279 \\ 0.310 \end{pmatrix}$$

$$\hat{\sigma^2} = S_{UU} - S_{UZ} S_{ZZ}^{-1} S_{ZU} = 0.737, \quad \hat{\sigma} = 0.858$$

2. 代入回归方程的绝对系数公式有:

$$R^2 = \frac{S_{UZ} S_{ZZ}^{-1} S_{ZU}}{S_{UU}} = 1 - \hat{\sigma^2} = 0.263$$

显著性检验统计量为:(注意这里一个自由度为约束个数 $q = 2$, 另一个为 $n - p = 20000 - 2 = 19998$, 中心化模型没有截距项，自变量个数和参数个数都是 $p = 2$).

$$F = \frac{n-p}{q} \times \frac{R^2}{1-R^2} \sim F_{q,n-p}$$

代入数值得

$$F = 3549.78$$

p 值计算为:

$$p = P\left(F_{2,19998} > 3549.78\right)$$

3. 在 $H_0 : \beta_1 = 0$ 下的 t 检验统计量为:

$$t = \frac{\hat{\beta}_1}{\hat{\sigma} / \left\| X^{\perp} \right\|} = \sqrt{n-p} \frac{r_p}{\sqrt{1 - r_p^2}} \sim t_{n-p}$$

其中 r_p 为 U 与 X 的偏相关系数, 计算为:

$$r_p = \frac{r_{UX} - r_{uv} r_{vx}}{\sqrt{1 - r_{uv}^2} \sqrt{1 - r_{vx}^2}} = \frac{0.438 - 0.453 \times 0.516}{\sqrt{1 - 0.453^2} \sqrt{1 - 0.516^2}} = 0.267$$

p 值为 $p = P\left(t_{19998} > 39.18\right)$

注 线性模型的 F 检验中，约束 $A_{q \times p} \beta_{p \times 1} = c_0$。$p$ 是模型参数的个数，在中心化模型里是自变量的个数，在一般模型里是自变量个数 +1 (多了一个截距参数)。这里 q 是约束的个数，是 A 的行数，A 的每一行给出一个约束方程。由于每一个约束相当于少了一个参数，所以 q 也可以表示零假设下的模型与原模型的参数个数之差。只有模型和课件上的完全相同才可以直接代公式。n 是样本量。

检验统计量

$$F = \frac{\left\| \hat{y} - \hat{y_0} \right\|^2 / q}{\left\| y - \hat{y} \right\|^2 / (n-p)}$$

可以理解为分子分母均除以各自的自由度：q 是 $\hat{y}$ 与 $\hat{y}_0$ 对应的模型参数个数之差，$n - p$ 是 y 与 $\hat{y}$ 对应的模型参数个数之差。

4.12 HW 12

作业 12 链接

✍ **练习 4.1** 假设 $(y_i, \mathbf{x}_i), i = 1, \ldots, n$ 满足线性模型

$$y_i = \beta_0 + \mathbf{x}_i^\top \mathbf{b} + \epsilon_i, \quad \epsilon_i \sim \left(0, \sigma^2\right), \quad i = 1, \ldots, n.$$

若对某个 $1 \le k \le n, y_k = \bar{y} = \sum_{i=1}^n y_i/n, \mathbf{x}_k = \bar{\mathbf{x}} = \sum_{i=1}^n \mathbf{x}_i/n$, 证明删除 $(y_k, \mathbf{x}_k)$ 不影响最小二乘法。

证明 线性模型可写为 $y = X\beta + \varepsilon$, $\hat{\beta}$ 为 β 的 LS 估计

设 $\widetilde{\beta}$ 为删除 $(\mathbf{x}_k, y_k)$ 后β 的 LS 估计, 根据 ppt week13 P29 命题 2

$$\begin{aligned}
\widetilde{\beta} = \hat{\beta} - \frac{\left(X^\top X\right)^{-1} \mathbf{x}_k e_k}{1 - h_{ii}} \\
\text{其中} e_k &= y_k - \mathbf{x}_k^\top \beta \\
&= \bar{y} - \bar{\mathbf{x}}^\top \beta \\
&= \frac{1}{n}\left(\mathbf{1}^\top \mathbf{y} - \mathbf{1}^\top \mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}\right) \\
&= \frac{1}{n} \mathbf{1}^\top \left(\mathbf{I} - \mathbf{P}_\mathbf{X}\right) \mathbf{y}
\end{aligned}$$

由于 $\mathbf{1} \in \mathbf{C}(\mathbf{X})$ ，则 $\mathbf{1}^\top \left(\mathbf{I} - \mathbf{P}_\mathbf{X}\right) = \mathbf{0}$ ，所以 $\tilde{\beta} = \hat{\beta}$

✍ **练习 4.2** 假设 $\mathbf{y}_1, \ldots, \mathbf{y}_n$ iid $\sim N\left(\theta, \sigma^2 I_p\right)$, 其中 θ 为 $p \times 1$ 未知参数向量, 假设 σ^2 已知。令 $\tilde{\theta} = \lambda \bar{\mathbf{y}}$, 其中 $0 \le \lambda \le 1$ 是常数。

(a) 求 $\tilde{\theta}$ 的均方误差 $m(\lambda) = E\|\tilde{\theta} - \theta\|^2$.

(b) 求 $\lambda_{opt} = \arg\min_\lambda m(\lambda)$ （即求使得 $m(\lambda)$ 达到最小的 λ).

(c) λ_{opt} 中含有未知 $\|\theta\|^2$, 以其无偏估计代入 (注意 $\|\bar{\mathbf{y}}\|^2$ 不是 $\|\theta\|^2$ 的无偏估计), 得到的 $\tilde{\theta} = \hat{\lambda}_{opt}\bar{\mathbf{y}}$ 是否等于或接近 James-Stein 估计?

解

(a) 因为

(i) $\mathrm{Var}(\tilde{\theta}) = \lambda^2 \mathrm{Var}(\bar{y}) = \lambda^2 \sigma^2 I_p/n$,

(ii) bias $= E\tilde{\theta} - \theta = \lambda\theta - \theta$

所以

$$m(\lambda) = E\|\tilde{\theta} - \theta\|^2 = \mathrm{tr}(\mathrm{Var}(\tilde{\theta})) + \|bias\|^2 = \lambda^2 \sigma^2 p/n + (\lambda - 1)^2 \|\theta\|^2$$

(b)

$$\lambda_{\text{opt}} = \underset{\lambda}{\operatorname{argmin}} m(\lambda) = \frac{\theta^\top \theta}{\frac{p\sigma^2}{n} + \theta^\top \theta}$$

(c)

$$\begin{aligned}
E\|\bar{y}\|^2 &= E\bar{y}^\top \bar{y} \\
&= \frac{np\sigma^2 + n^2 \theta^\top \theta}{n^2} \\
&= \frac{p\sigma^2}{n} + \theta^\top \theta
\end{aligned}$$

$$\|\bar{y}\|^2 - \frac{p\sigma^2}{n} \text{为} \theta^\top \theta \text{的无偏估计}$$

$$\begin{aligned}
\tilde{\theta} = \hat{\lambda}_{\text{opt}}\, \bar{y} &= \frac{\|\bar{y}\|^2 - \frac{p\sigma^2}{n}}{\|\bar{y}\|^2} \bar{y} \\
&= \left(1 - \frac{p\sigma^2}{n\|\bar{y}\|^2}\right) \bar{y}
\end{aligned}$$

$$\text{James - Stein 估计}\hat{\theta}_{J-S}=\left(1-\frac{(p-2)\sigma^2}{n\|\bar{y}\|^2}\right)\bar{y}$$

✍ **练习 4.3** 假设模型

$$\mathbf{y}_{n\times 1}=X_{n\times p}\beta_{p\times 1}+\epsilon_{n\times 1},\epsilon\sim\left(0,\sigma^2 I_n\right),$$

其中 X 的第一列为向量 $\mathbf{1}$. 回归系数的最小二乘估计记为 $\hat{\beta}=\left(X^\top X\right)^{-1}X^\top\mathbf{y}$ 。假设有“新”数据 $\mathbf{x}_0,y_0$ 满足模型 $y_0=\mathbf{x}_0^\top\beta+\epsilon_0,\epsilon_0\sim\left(0,\sigma^2\right)$, 其中 $\mathbf{x}_0$ 已知, 需要预测 y_0 。预测统计量取为 $\hat{y}_0=\mathbf{x}_0^\top\hat{\boldsymbol{\beta}}$ 证明如下结论

(a) $\hat{y}_0$ 的预测误差为 $pe\left(\hat{y}_0\right)=E\left(\hat{y}_0-y_0\right)^2=\sigma^2\left[1+\mathbf{x}_0^\top\left(X^\top X\right)^{-1}\mathbf{x}_0\right]$.

(b) 当 $\mathbf{x}_0=\mathbf{x}_i$($X$ 的第 i 行), $pe\left(\hat{y}_0\right)=\left(1+h_{ii}\right)\sigma^2\overset{\text{记作}}{=}e\left(\mathbf{x}_i\right)$, 其中 h_{ii} 为 $H=X\left(X^\top X\right)^{-1}X^\top$ 的 (i,i) 元;

(c) 当 $\mathbf{x}_0=\bar{\mathbf{x}},\hat{y}_0=\bar{y}=\mathbf{1}^\top\mathbf{y}/n$, 且 $pe\left(\hat{y}_0\right)=(1+1/n)\sigma^2\overset{\text{记作}}{=}e(\bar{\mathbf{x}})$, 说明 $e(\bar{\mathbf{x}})\leq e\left(\mathbf{x}_i\right),i=1,\ldots,n$.

(d) 当 $\mathbf{x}_0=X^\top\mathbf{a}$ (这里 $\mathbf{a}\in R^n$) , $pe\left(\hat{y}_0\right)\leq\sigma^2\left(1+\|\mathbf{a}\|^2\right)$.

证明

(a)

$$\begin{aligned}
\mathrm{pe}\left(\hat{y}_0\right)&=E\left(\hat{y}_0-y_0\right)^2\\
&=E\left(\mathbf{x}_0^\top\hat{\beta}-\mathbf{x}_0^\top\beta-\varepsilon_0\right)^2\\
&=E\left(\mathbf{x}_0^\top\hat{\beta}-\mathbf{x}_0^\top\beta\right)^2+E\varepsilon_0^2\\
&=\mathrm{Var}\left(\mathbf{x}_0^\top\hat{\beta}\right)+\sigma^2\\
&=\mathbf{x}_0^\top\mathrm{Var}(\hat{\beta})\mathbf{x}_0+\sigma^2\\
&=\mathbf{x}_0^\top\mathrm{Var}\left(\left(X^\top X\right)^{-1}X^\top y\right)\mathbf{x}_0+\sigma^2\\
&=\left(1+\mathbf{x}_0^\top\left(X^\top X\right)^{-1}\mathbf{x}_0\right)\sigma^2
\end{aligned}$$

(b) 当 $\mathbf{x}_0=\mathbf{x}_i$ 时,

$$\begin{aligned}
&\mathbf{x}_i\left(X^\top X\right)^{-1}\mathbf{x}_i=h_{ii}\\
\Rightarrow&pe\left(\hat{y}_0\right)=\left(1+h_{ii}\right)\sigma^2
\end{aligned}$$

(c) 当 $\mathbf{x}_0=\bar{\mathbf{x}}$ 时

$$\begin{aligned}
\bar{\mathbf{x}}^\top\left(X^\top X\right)^{-1}\bar{\mathbf{x}}&=\frac{1}{n^2}\mathbf{1}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{1}\\
&=\frac{1}{n^2}\mathbf{1}^\top\mathbf{P_X}\mathbf{1}\\
&=\frac{1}{n^2}\mathbf{1}^\top\mathbf{1}\\
&=\frac{1}{n}\\
\mathrm{pe}\left(\hat{y}_0\right)&=\left(1+\frac{1}{n}\right)\sigma^2
\end{aligned}$$

$$\text{根据 ppt weck 13 命题 1 可知}h_{ii}\geq\frac{1}{n}$$

所以 $e(\bar{\mathbf{x}})\leqslant e\left(\mathbf{x}_i\right)$

(d) 当 $\mathbf{x}_0=X^\top a$ 时

$$\begin{aligned}
pe\left(\hat{y}_0\right)&=\left(a^\top P_X a+1\right)\sigma^2\\
&=\left(\|P_X a\|^2+1\right)\sigma^2\\
&\leqslant\left(\|a\|^2+1\right)\sigma^2
\end{aligned}$$

✍ **练习 4.4** 条件同上一题, 对任何给定的 $\mathbf{x}_0\in R^p$, 我们需要预测对应的 y_0, 但 y_0 的预测取为 $\tilde{y}_0=\mathbf{x}_0^\top\tilde{\boldsymbol{\beta}}$, 其中 $\tilde{\beta}=\hat{\beta}\lambda,0\leq\lambda\leq 1$ 。

(a) 证明 $\tilde{y}_0$ 的预测误差为

$$\mathrm{pe}\left(\tilde{y}_0\right)=E\left(\tilde{y}_0-y_0\right)^2=(1-\lambda)^2\left(\mathbf{x}_0^\top \beta\right)^2+\lambda^2 \sigma^2 \mathbf{x}_0^\top\left(X^\top X\right)^{-1} \mathbf{x}_0+\sigma^2 .$$

(b) 证明如果 $\|X \beta\|^2 \leq \frac{1+\lambda}{1-\lambda} \sigma^2$, 则 $pe\left(\tilde{y}_0\right) \leq pe\left(\hat{y}_0\right)$, 其中 $\hat{y}_0=\mathbf{x}_0^\top \hat{\boldsymbol{\beta}}$ 为基于 LS 估计的预测。

证明

(a)

$$\begin{aligned}
\tilde{y}_0 &= \mathbf{x}_0^\top \tilde{\beta}=\lambda \mathbf{x}_0^\top \hat{\beta} \\
p\left(\tilde{y}_0\right) &= E\left(\tilde{y}_0-y_0\right)^2 \\
&= E\left(\lambda \mathbf{x}_0^\top \hat{\beta}-\mathbf{x}_0^\top \beta-\varepsilon_0\right)^2 \\
&= E\left(\lambda \mathbf{x}_0^\top \hat{\beta}-\lambda \mathbf{x}_0^\top \beta+(\lambda-1) \mathbf{x}_0^\top \beta-\varepsilon_0\right)^2 \\
&= \operatorname{Var}\left(\lambda \mathbf{x}_0^\top \hat{\beta}\right)+(1-\lambda)^2\left(\mathbf{x}_0^\top \beta\right)^2+E \varepsilon_0^2 \\
&= \lambda^2 \mathbf{x}_0^\top \operatorname{Var}(\hat{\beta}) \mathbf{x}_0+(1-\lambda)^2\left(\mathbf{x}_0^\top \beta\right)^2+\sigma^2 \\
&= \lambda^2 \mathbf{x}_0^\top\left(X^\top X\right)^{-1} \mathbf{x}_0 \sigma^2+(1-\lambda)^2\left(\mathbf{x}_0^\top \beta\right)^2+\sigma^2
\end{aligned}$$

(b)

$$\begin{aligned}
\mathrm{pe}\left(\tilde{y}_0\right) &= (1-\lambda)^2\left(\mathbf{x}_0^\top \beta\right)^2+\lambda^2 \mathbf{x}_0^\top\left(X^\top X\right)^{-1} \mathbf{x}_0 \sigma^2+\sigma^2 \\
pe\left(\hat{y}_0\right) &= \mathbf{x}_0^\top\left(X^\top X\right)^{-1} \mathbf{x}_0 \sigma^2+\sigma^2 \\
\mathrm{pe}\left(\hat{y}_0\right)-\mathrm{pe}\left(\tilde{y}_0\right) &= \left(1-\lambda^2\right) \mathbf{x}_0^\top\left(X^\top X\right)^{-1} \mathbf{x}_0 \sigma^2-(1-\lambda)^2\left(\mathbf{x}_0^\top \beta\right)^2 .
\end{aligned}$$

$$\text{由于} \beta^\top X^\top X \beta \leqslant \frac{1+\lambda}{1-\lambda} \sigma^2 \text{和 ppt week15 p5 引理 1}$$

$$\begin{aligned}
&\beta \beta^\top \leqslant \frac{1+\lambda}{1-\lambda} \sigma^2\left(X^\top X\right)^{-1} \\
&\mathbf{x}_0^\top \beta \beta^\top \mathbf{x}_0 \leqslant \frac{1+\lambda}{1-\lambda} \mathbf{x}_0^\top\left(X^\top X\right)^{-1} \mathbf{x}_0 \sigma^2 \\
&\text{所以} pe\left(\hat{y}_0\right) \geqslant pe\left(\tilde{y}_0\right)
\end{aligned}$$

练习 4.5 假设模型 $\mathbf{y}_{n \times 1}=X_{n \times p} \beta_{p \times 1}+\epsilon_{n \times 1}, \epsilon \sim\left(0, \sigma^2 I_n\right)$ 中 $X, \mathbf{y}$ 都已经中心化 (因此模型中没有截距项), 对设计阵 X 进行奇异值分解:

$$X_{n \times p}=U_{n \times p} D_{p \times p} V_{p \times p}^\top,$$

其中 $U^\top U=I_p, V^\top V=I_p, D=\operatorname{diag}\left(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_p}\right)$, 这里 $\lambda_1 \geq \ldots \geq \lambda_p>0$ 是 $X^\top X$ 的特征根。所以 $X \beta=U D V^\top \beta=U \gamma$, 其中 $\gamma=D V^\top \beta$ 。由此, 我们改写原模型为

$$\mathbf{y}=U \gamma+\epsilon, \epsilon \sim\left(0, \sigma^2 I_n\right),$$

此模型称为主成分回归模型。记 U 的各列为 $U=\left(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p\right)$ 。

(a) 证明主成分回归模型中回归系数 γ 的最小二乘估计为 $\hat{\gamma}=U^\top \mathbf{y}$, 由此求出原模型中的 β 的 LS 估计。

(b) 主成分回归模型中, 响应变量的基于最小二乘估计 $\hat{\gamma}$ 的拟合值向量为 $\hat{\mathbf{y}}=U \hat{\gamma}=\sum_{j=1}^p \mathbf{u}_j\left(\mathbf{u}_j^\top \mathbf{y}\right)$, 求其均方误差 $m(\hat{\mathbf{y}})=E\|\hat{\mathbf{y}}-U \gamma\|^2$.

(c) 设下标集合 $A_q=\left\{i_1, \ldots, i_q\right\} \subset\{1,2, \ldots, p\}, 1 \leq q \leq p-1$, 令

$$\tilde{\mathbf{y}}^{\left(A_q\right)}=\sum_{j \in A_q} \mathbf{u}_j\left(\mathbf{u}_j^\top \mathbf{y}\right) .$$

求其均方误差 $m\left(\tilde{\mathbf{y}}^{\left(A_q\right)}\right)=E\left\|\tilde{\mathbf{y}}^{\left(A_q\right)}-U \gamma\right\|^2$ 。假设 $\|\gamma\|^2 \leq \sigma^2$, 证明

$$m\left(\tilde{\mathbf{y}}^{\left(A_q\right)}\right) \leq m(\hat{\mathbf{y}}).$$

(d) 写出 C_q 准则的具体表达, 给出最优子集搜索算法。

解

(a)

$$
\begin{aligned}
\hat{\gamma} &= \left(U^\top U\right)^{-1} U^\top y \\
&= U^\top y \\
&= DV^\top \beta \\
X\beta &= U\hat{\gamma} \\
&= UU^\top y \\
X\left(X^\top X\right)^{-1} X^\top &= UDV^\top \left(VDU^\top UDV^\top\right)^{-1} VDU^\top \\
&= UU^\top
\end{aligned}
$$

$$
\begin{aligned}
\Rightarrow X\beta &= X\left(X^\top X\right)^{-1} X^\top y \\
\hat{\beta} &= \left(X^\top X\right)^{-1} X^\top y
\end{aligned}
$$

(b)

$$
\begin{aligned}
m(\hat{y}) &= E\|\hat{y} - U\gamma\|^2 \\
&= E\left\|UU^\top y - U\gamma\right\|^2 \\
&= E\left\|UU^\top (y - U\gamma)\right\|^2 \\
&= E\left\|UU^\top \varepsilon\right\|^2 \\
&= E\left(\varepsilon^\top UU^\top \varepsilon\right) \\
&= \operatorname{tr}\left(UU^\top\right)\sigma^2 \\
&= \operatorname{tr}(U^\top U)\sigma^2 \\
&= p\sigma^2
\end{aligned}
$$

(c)

$$
\begin{aligned}
\tilde{y}^{(A_q)} &= \sum_{j\in A_q} u_j\left(u_j^\top y\right) \\
m\left(\tilde{y}^{(A_q)}\right) &= E\left\|\sum_{j\in A_q} u_j\left(u_j^\top y\right) - \sum_{i=1}^{p} u_j\gamma_i\right\|^2 \\
&= E\left\|\sum_{j\in A_q} u_j\hat{\gamma}_j - \sum_{i=1}^{p} u_i\gamma_i\right\|^2 \\
&= E\left\|\sum_{j\in A_q} u_j\left(\hat{\gamma}_j-\gamma_j\right) - \sum_{i\notin A_q} u_i\gamma_i\right\|^2 \\
&= E\left\|\sum_{j\in A_q} u_j\left(\hat{\gamma}_j-\gamma_j\right)\right\|^2 + E\left\|\sum_{i\notin A_q} u_i\gamma_i\right\|^2 \\
&= E\left(\sum_{j\in A_q} u_j\left(\hat{\gamma}_j-\gamma_j\right)\right)^\top\left(\sum_{j\in Aq} u_j\left(\hat{\gamma}_j-\gamma_j\right)\right) + \sum_{i\notin A_q}\gamma_i^2 \\
&= \sum_{j\in A_q} E\left\|\hat{\gamma}_j-\gamma_j\right\|^2 + \sum_{i\notin A_q}\gamma_i^2 \\
&= \sum_{j\in A_q} E\varepsilon^\top u_j u_j^\top\varepsilon + \sum_{i\notin A_q}\gamma_i^2 \\
&\leqslant q\sigma^2 + \|\gamma\|^2 \\
&\leqslant (q+1)\sigma^2 \leqslant p\sigma^2 = m(\hat{y})
\end{aligned}
$$

(d)

$$
\begin{aligned}
C_q &= \frac{RSS_q}{\hat{\sigma}^2} - n + 2q \\
\text{其中}\hat{\sigma}^2 &= \frac{RSS_p}{x-p} \\
RSS_q &= \left\|y - \tilde{y}^{(A_q)}\right\|^2 \\
&= \|y\|^2 - \sum_{j\in A_q}\left(u_j^\top y\right)^2 \\
RSS_p &= \|y-\hat{y}\|^2 \\
&= \|y\|^2 - \sum_{j=1}^{p}\left(u_j^\top y\right)^2
\end{aligned}
$$

$$
C_q = \frac{\|y\|^2 - \sum_{j\in A_q}\left(u_j^\top y\right)^2}{\|y\|^2 - \sum_{j=1}^{p}\left(u_j^\top y\right)^2}(n-p) - n + 2q
$$

1. 计算 $u_j^\top y, j=1,\ldots,p$, 并排序 $u_{i_1}^\top y \leqslant \cdots \leqslant u_{i_p}^\top y$

2.$q^* = \underset{q}{\operatorname{argmin}} = \frac{\|y\|^2 - \sum_{j\in A_q}\left(u_j^\top y\right)^2}{\hat{\sigma}^2}(n-p) - n + 2q$

3.$A_{q^*} = \{i_1,\ldots,i_{q^*}\}$

第 5 章 复习

本章内容直接来源于助教曾经的课程整理，这是思维导图的链接。而后者直接来源于杨老师的课件。
注 我们旨在梳理课程内容，建立一个完整的知识框架，可用于查漏补缺，但不代表考试重点，如有错误疏漏请指正。

5.1 因果推断

因果：其他变量给定下的本质关系，判断因果（找干扰因素、是否因果颠倒、前提：条件均同）。
关联：不考虑其他因素得到的表面关系，与关联对立的概念是独立。
研究设计：自变量取值是否由外界干预。随机化控制试验（因果推断所必须）、观察研究。

5.2 相关性

5.2.1 相关系数和偏相关系数

相关系数：定义。区分总体相关系数 ρ 和样本相关系数 r.
偏相关系数：三元（具体形式可写出, 建议记忆）、多元（矩阵表示，代表去相关化的向量的相关系数）。$\Sigma_{ab\bullet c} = \Sigma_{ab} - \Sigma_{ac}\Sigma_{cc}^{-1}\Sigma_{cb}$.

5.2.2 相关性检验

精确检验：正态假设下相关系数的分布, 熟悉检验统计量是 t 分布的证明。

$$t = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}} \sim t_{n-2}. \tag{5.1}$$

大样本检验

$$z = \sqrt{n} \times r. \tag{5.2}$$

置换检验: 为什么可以得到 p 值?
和两样本 t 检验与 Pearson 卡方检验的关系（熟悉课件对应内容、作业题的证明）

- 正态假设下，检验 r 和 t 相同。
- 二项假设下，检验 r 和卡方形式类似。

5.2.3 辛普森悖论

总体指标与单项指标之间的差异情况不一致，蛮有趣。

5.2.4 随机向量

随机向量的方差、协方差矩阵的定义。
公式

$$E(x^\top Cx) = \mu^\top C\mu + tr(\Sigma C), \tag{5.3}$$

其中 $Ex = \mu, var(x) = \Sigma$.
相关系数矩阵（即归一化的协方差矩阵）
去相关化（协方差阵为 0）
偏相关系数矩阵（涉及分块矩阵的求逆公式）

两个随机向量整体的相关性：正交分解、方差分解、典则相关系数（决定系数、方差被解释的比例）、条件整体相关性。

注 这里的决定系数和后面线性模型中的是一样的。线性模型也可以认为是在研究 y 和向量 x 整体的相关性。在学完整门课以后复习到这里，应该可以领会在“回归分析”的前几节只讲“相关性”的意义所在。

5.3 线性回归模型

5.3.1 总体线性模型

理解总体模型。总体模型和随机向量的去相关化是一样的。

Gauss-Markov 假设。

参数矩估计。（除了误差方差的估计，和最小二乘估计相同）

控制变量的线性模型。（实质是随机向量的偏相关研究）

5.3.2 单变量简单线性模型

最小二乘法：正则方程（非常重要!）、性质、LS 估计的均值和方差、平方和、样本版本的决定系数、误差方差的最小二乘估计（无偏性）。

RSS 可以用误差表示：

$$RSS = s_{yy} - s_{xy}^2/s_{xx} = s_{\varepsilon\varepsilon} - s_{x\varepsilon}^2/s_{xx}. \tag{5.4}$$

Wald 检验：熟悉如何推导这个 t 统计量的分布。

回归效应：“回归”名字的由来。

5.3.3 多元回归

投影：定义，对称幂等线性变换，最优性（投影提供的最小二乘估计是残差平方和最小的估计），投影阵。

多变量线性回归模型

- 最小二乘法：拟合值、残差、误差方差的估计、LS 估计无偏性和方差、LS 估计的最优性（Gauss-Markov 定理）、分量的估计（去相关化）
- 拟合优度（得到的模型拟合效果是否好?）：方差正交分解、样本版本的决定系数（是样本相关系数的平方，即 $R^2 = r^2$）.
- 中心化：一元、多元。熟悉诸如 $Z_c^\top Z_c = Z^\top(I_n - P_1)Z$ 的表达。线性模型的中心化不改变回归系数 b，只改变截距项，简化计算。
- 约束最小二乘和 F 检验：本节中所有内容都可以写成线性模型最小二乘法，加一个约束 $A\beta = c_0$（c_0 为什么可以不妨为 0?）的形式。
  - 约束最小二乘的求解：矩阵求导
  - 几何解释：最小二乘本质是投影，约束最小二乘便是在子空间中投影。在这一节证明中关键的引理便是在求子空间 V_0 的投影阵 P_{V_0}.
  - 一般线性假设是什么?
  - Wald 检验
  - F 检验的不同表示形式。在原假设成立的前提下，

    $$F = \frac{n-p}{q} \times \frac{\|\hat{y} - \hat{y}_0\|^2}{\|y - \hat{y}\|^2} \sim F_{q,n-p}. \tag{5.5}$$

    实际问题中通常可以具体得到 V_0，从而简化计算。这部分参考对应内容的作业解答，那里对于问题的切入点写得比较详细。

    检验部分回归系数是否为 0，两个特例是回归方程的显著性检验和单个回归系数的 t 检验。

- 广义最小二乘。定义（误差加权），最优线性无偏估计（BLUE），与 OLS 的方差比较，与多元正态对应的极大似然比较。
  加权最小二乘：方差更小，计算方法，迭代加权最小二乘（IRLS）.

5.4 方差分析

注 本节最核心的内容是把具体问题写成线性模型的矩阵形式。方差分析可以视为线性模型的应用，其理论和前面的多元回归理论是相同的。

5.4.1 单因素方差分析

线性模型形式，F 检验（anv 方差分析表，特例：两样本 t 检验），对比（重新参数化），两样本 t 检验（成对，二值 binary 数据，参见作业）

5.4.2 两因素方差分析

可加性假设，可加方差分析模型，F 检验（配对设计）。
注 梳理下来发现，我们在学习多元模型的时候，总去算很多二元或三元的例子（课件、作业题）。这是很好地练习，在计算量不大的同时，也让回归分析的理论触手可及。单纯证明抽象的高维问题虽然从理论上来讲足够，但是也会让理论丧失直观的理解。

5.5 回归诊断

5.5.1 残差分析

对模型假设（线性性和方差齐性）的合理性进行诊断
残差图可以看什么? 线性、方差齐性
Box-Cox 变换

5.5.2 影响分析

影响分析是为了发现对回归结果影响较大的点。

- 异常点: 定义（y 值异常，残差较大），标准化残差，学生化残差。何种程度可以视为异常?
- 高杠杆点：定义（x 值异常），判别标准
- 高影响点（未必异常），Cook 距离 D，删掉一个点对于模型的影响（理论分析），Cook 距离的定义和计算。

5.5.3 6 个图

残差图，qq-nrom，刻度-位置图，Cook's D, 残差-杠杆图，D vs h(影响分析).

5.6 预测

5.6.1 预测误差分解

预测误差分解，均方误差分解，James-Stein 估计，矩阵形式如何写。

5.6.2 线性模型的预测

基于子模型的预测，预测误差。

5.6.3 变量选择

Cp 准则 (熟悉理论，如何操作？)，AIC 与 BIC.

5.6.4 最优子集回归

计算复杂，设计阵列正交时可简化计算，主成分回归。

5.6.5 变量选择的贪心算法

5.7 写在最后

建议复习的方法是根据老师在课程最后给出的考试范围，把范围内的课件通读一遍，保证在读后至少可以在再次翻到某部分的时候，顺利地看懂。具体的数学细节（证明、推导）需要理解、掌握、记忆到什么程度自己把握。另外，很多重要的公式、结论建议记清楚，统计课毕竟是统计课。