Maxwell General Learning System Academic Statistical Modeling
麦克斯韦通用高等习得系统 统计建模

Leo_Maxwell

March 2023

目录

序言 Introduction

本书主要讲了如何通过运用统计方法建立数学模型并进行分析。
您至少需要掌握《综合统计学》中大部分的初等统计学知识才能使用这份文档。

第一章　统计量推广 Generalization

1.1　随机矩阵 Random Matrices

定义：若一个向量或矩阵中的每一个元素都是一个随机变量，则称该向量或矩阵为随机向量或随机矩阵。
下文如无特殊说明不再区分随机矩阵和随机向量。

1.2　期望 (Mean)

对随机矩阵 $\boldsymbol{X}$ 而言，其期望是一个尺寸相同的矩阵，并且对应位置的元素为该元素的期望值。
容易证明随机矩阵的期望满足一般期望的所有性质。

1.3　方差 (Variance)

此处的方差仅针对随机向量定义。对随机向量 $\overrightarrow{X} = (X_1, X_2, X_3, \ldots, X_n)$，定义其方差 $\operatorname{Var}(()\overrightarrow{X})$ 为一 $n \times n$ 的矩阵 Σ，其中第 i 行第 j 列的元素 $\Sigma_{ij} = \operatorname{Cov}(()X_i, X_j)$.

它和一般的方差性质的区别主要在以下这一条：对长度为 n，方差为 $\boldsymbol{\Sigma}$ 的随机向量 $\overrightarrow{X}$（列向量）和尺寸为 $m \times n$ 的矩阵 $\boldsymbol{A}$ 和与 $\overrightarrow{X}$ 同尺寸的向量 $\overrightarrow{b}$，有：（证明见7.1.1）

$$\operatorname{Var}(()\boldsymbol{A}\overrightarrow{X} + \overrightarrow{b}) = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T$$

1.4　分布表示 Distribution Notation

对一个长度为 n 的随机向量 $\overrightarrow{X}$ 而言，如果它符合期望是 $\overrightarrow{\mu}$，方差是矩阵 $\boldsymbol{\Sigma}$ 的 n 元正态分布，则记：

$$\overrightarrow{X} \sim N_n(\overrightarrow{\mu}, \boldsymbol{\Sigma})$$

第二章　回归分析 (Regression Analysis)

对存在一定关系的两个随机变量而言，我们希望用一个代数意义上的函数来描述这两个变量之间的关系。

由于它们都是随机变量，所以天然地存在不确定性，不可能存在传统意义上的 $y = f(x)$ 的数学关系。然而，如果我们能设法得到 $\mathrm{E}[Y] = \mathrm{E}[f(X)]$ 这样的关系，即使用随机变量的期望作为其主要取值，然后得到函数关系，由于现实世界里得到的绝大多数数据本质上都是随机变量，所以获得这种关系的努力有显著的现实意义。

例如，一般来说体重越重的人身高也越高，但很显然，在人群中，这两个量都是随机变量。它们存在什么样的关系呢？我们希望用一个上述的函数来描述这种关系。

2.1　定义

对 $(x_{11}, x_{12}, x_{13}, \ldots, x_{1p}, y_1), (x_{21}, x_{22}, x_{23}, \ldots, x_{2p}, y_2), (x_{31}, x_{32}, x_{33}, \ldots, x_{3p}, y_3), \ldots, (x_{n1}, x_{n2}, x_{n3}, \ldots, x_{np}, y_n)$ 而言，我们假设存在一组 $\beta_j, j = 0, 1, 2, \ldots, p$，使得 y_i 能够被以下的式子表达：

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \epsilon_i$$

并且成立以下条件：

残差的期望为零： $\mathrm{E}[\overrightarrow{\epsilon}] = \overrightarrow{0}$

每个残差的方差都相同，且互不相关： $\mathrm{Var}(\overrightarrow{\epsilon}) = \sigma^2 \boldsymbol{I}$

或是一个更严格的条件：

$$\overrightarrow{\epsilon} \sim N(\overrightarrow{0}, \sigma^2 \boldsymbol{I})$$

则称这些公式和数字构成一个**简单线性回归模型 (Simple Linear Regression Model, SLR Model)**。

其中，所有 x 被称为**自变量**，所有 y 被称为**因变量**，所有 ϵ 被称为**残差**。于是该定义是直观的：它试图通过已经获得的大量数据推知许多个 β，然后再通过这些 β 和已知的自变量推知因变量。

容易知道这种关系可以通过矩阵表示：

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & x_{23} & \cdots & x_{2p} \\ 1 & x_{31} & x_{32} & x_{33} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \cdots & x_{np} \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

化简为

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

即得到简单线性回归模型的一般表示。其中，矩阵 $\boldsymbol{X}$ 被称作是**设计矩阵 (Design Matrix)**。

该定义通过引入残差将随机性从自变量和因变量中去除了，即残差本身概括了自变量和因变量的所有随机特征。这种叙述的数学表达就是，$\overrightarrow{y}$ 的方差事实上就是残差的方差：

$$\mathrm{Var}\left(\overrightarrow{y}\right) = \mathrm{Var}\left(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\epsilon}\right)$$

当我们用后续的各种办法求出 $\boldsymbol{\beta}$ 的估计值以后，我们可以将估计值作为 $\boldsymbol{\beta}$ 的取值，于是 $\boldsymbol{\beta}$ 和 $\boldsymbol{X}$ 都成为定值，所以得到 $\mathrm{Var}\left(\overrightarrow{y}\right) = \mathrm{Var}\left(\boldsymbol{\epsilon}\right)$.

在该定义中需要注意 $\boldsymbol{\beta}$ 的属性。此处它仅代表一个在符合假设的情形下的一个未知的矩阵，它不是一个随机变量。

2.2 最小二乘法 (Least Square Estimation, LSE)

使得下列多元函数取得最小值的一组 β，就是用最小二乘法得出的理想的 $\boldsymbol{\beta}$，我们记为 $\hat{\boldsymbol{\beta}}$。

$$Q(\beta_0, \beta_1, \beta_2, \ldots, \beta_p) = \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j \right) \right)^2$$

所以其得名最小二乘法。

该函数可使用线性代数表示，为 $\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$，此处定义的范数是 Euclidean 范数。

需要注意此处 $\hat{\boldsymbol{\beta}}$ 的属性。由于 $\hat{\boldsymbol{\beta}}$ 必须满足令残差平方和最小的定义，所以它是一个依赖于随机变量的变量，即随机变量。

最小二乘法要求设计矩阵的列是满秩的。否则，下文计算该矩阵的方法便不能使用，这种要求的直观理解是：若设计矩阵的列不满秩，则不线性独立的列所代表的预测变量没有意义，因为该预测变量能够被其他预测变量所完全表示。这意味着最终计算得到的系数矩阵不是唯一的。从另一种角度而言，也可以认为不线性独立的列所代表的预测变量不提供任何新的有价值的信息，所以在实际建模的操作中通常将这种预测变量剔除。

2.2.1 计算系数矩阵 Calculating the LSE Coefficients

$\hat{\boldsymbol{\beta}}$ 可通过以下方式计算：（证明请见7.1.3）

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

2.2.2 拟合值 Fitted Values

根据最小二乘法的定义，得知拟合的 y 值是：

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

我们记矩阵 $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$，得到：

$$\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$$

矩阵 $\boldsymbol{H}$ 是 n 阶方阵。它有一些很好的性质，由其定义计算可立得：

- $\boldsymbol{H} = \boldsymbol{H}^T$
- $\boldsymbol{H}^n = \boldsymbol{H} \quad (n \geq 1 \in \mathbb{Z})$
- $(\boldsymbol{I} - \boldsymbol{H})^n = \boldsymbol{I} - \boldsymbol{H} \quad (n \geq 1 \in \mathbb{Z})$

2.2.3 系数矩阵的统计性质 Statistical Properties of LSE Coefficients

前文提到，最小二乘的系数矩阵是随机变量。于是它有统计性质：（证明见7.1.5）

- $\mathrm{E}\left[\hat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta}$
- $\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$
- $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$

2.2.4 估计方差 Estimating the Variance

在最小二乘法中，使用该方法估计残差的方差 σ^2：（证明请见7.1.4）

$$s_e^2 = \frac{1}{n-p-1}\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2$$

该方差的无偏估计有类似于其他方差估计的性质：（证明请见7.1.4）

$$\frac{(n-p-1)s_e^2}{\sigma^2} \sim \chi^2(n-p-1)$$

2.2.5 Gauss-Markov Theorem

对于 $\boldsymbol{\beta}$ 而言，它的任意一个线性函数 $\lambda(\boldsymbol{\beta})$ 的最小方差线性无偏估计 (MVLUE) 是 $\lambda(\hat{\boldsymbol{\beta}})$，其中 $\hat{\boldsymbol{\beta}}$ 是最小二乘法得到的估计。该定理的证明请见7.1.6。

2.2.6 最小二乘中的假设检验 Hypothesis Test in LSE

2.2.6.1 枢轴量 Pivotal Quantity

对 $\boldsymbol{\beta}$ 中第 i 个元素 β_i 而言，可以构造枢轴量如下：

$$\frac{\hat{\beta}_i - \beta_i}{s_e\sqrt{c_{ii}}}$$

其中，c_{ii} 是矩阵 $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ 中的第 i 行第 i 列的元素，即对角线上的元素。该枢轴量有性质：

$$\frac{\hat{\beta}_i - \beta_i}{s_e\sqrt{c_{ii}}} \sim t(n-p-1)$$

2.3 岭回归 (Ridge Regression)

使得下列多元函数取得最小值的一组 β，就是用岭回归得出的理想的 $\boldsymbol{\beta}$，我们记为 $\hat{\boldsymbol{\beta}}$.

第三章　分类模型 Classification Models

分类模型的目标是根据给定的多个对象的数据将这些对象根据某个标准分成 k 类。分类模型可以根据已有的类型数据生成分类标准，然后对新的对象进行分类，这一般被称作**监督学习 (Supervised Learning)**。或者，分类模型还可以在没有已知分类数据的情况下，将多个对象分成几类，相对地，这被称为**无监督学习 (Unsupervised Learning)**。

3.1 朴素 Bayes 分类 Naive Bayes Classifier

3.1.1 Bayes 定理 Bayes Theorem

对事件 A 和事件 B 而言，条件概率 $P(A|B)$ 满足：

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

它的证明是平凡的：根据条件概率定义立得 $P(A|B) = \dfrac{P(AB)}{P(B)}, P(B|A) = \dfrac{P(AB)}{P(A)}$，代入后得到 Bayes 定理。

3.1.2 假设 Assumption

朴素 Bayes 要求用于分类的多个特征之间相互独立。

3.1.3 建立模型 Model Construction

我们有一组已知的分类数据，即矩阵 $\boldsymbol{X}$：

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2n} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & x_{m3} & \cdots & x_{mn} \end{pmatrix}$$

其中，已知矩阵中每一行特征对应的对象所属的分类，并且一共有 K 类，m 个对象。用 k_i 表示属于第 i 类的对象的个数，并令 $\pi_i = k_i/m$，将它称之为分类 i 的**先验概率 (A Priori)**。

现在给定一个待分类的对象 y 的特征 $\overrightarrow{y} = (y_1, y_2, y_3, \ldots, y_n)$。我们用 Y_k 表示对象 y 属于第 k 类的这个事件，用 $\boldsymbol{X}$ 表示获得了这一组分类信息的这一事件。那么根据 Bayes 定理，我们得到：

$$P(Y_k|\boldsymbol{X}, \overrightarrow{y}) = \frac{P(\boldsymbol{X}, \overrightarrow{y}|Y_k)P(Y_k)}{P(\boldsymbol{X}, \overrightarrow{y})}$$

则朴素 Bayes 分类的思想就是找到使得该条件概率最大的分类 k，然后将对象 y 分类到这一类中去。接下来对上式进行变形，使得它能够被计算。

注意到上式的分母在给定这一组分类数据后是常量，所以该条件概率的大小完全取决于分子的大小，得到：

$$P(Y_k|\boldsymbol{X},\overrightarrow{y}) \propto P(\boldsymbol{X},\overrightarrow{y}|Y_k)P(Y_k)$$

实际上，上式右侧就是联合概率 $P(Y_k,\boldsymbol{X},\overrightarrow{y})$，并且我们认为此处的 $\boldsymbol{X}$ 提供了 $\overrightarrow{y}$ 中每一个特征 y_i 的分布信息。这里的分布信息指的主要是属于第 k 类的所有对象的第 i 个特征组成的样本的分布信息。我们在后文中还会用到此定义。

接下来将 $\boldsymbol{X}$ 从式子中暂时去除（但该式仍然依赖 $\boldsymbol{X}$ 所含的信息），并写成这种形式：

$$P(Y_k|\boldsymbol{X},\overrightarrow{y}) \propto P(Y_k,y_1,y_2,y_3,\cdots,y_n)$$

利用链式法则，写成：

$$\begin{aligned}P(Y_k|\boldsymbol{X},\overrightarrow{y}) &\propto P(Y_k,y_1,y_2,y_3,\cdots,y_n)\\ &\propto P(Y_k)P(y_1,y_2,y_3,\cdots,y_n|Y_k)\\ &\propto P(Y_k)P(y_1|Y_k)P(y_2,y_3,y_4,\cdots,y_n|Y_k,y_1)\\ &\propto P(Y_k)P(y_1|Y_k)P(y_2|Y_k,y_1)P(y_3,y_4,y_5,\cdots,y_n|Y_k,y_1,y_2)\\ &\cdots\end{aligned}$$

利用所有特征之间相互独立的假设，得到 $P(y_i|Y_k,y_j)=P(y_i|Y_k)$，于是上式可以写成：

$$\begin{aligned}P(Y_k,y_1,y_2,y_3,\cdots,y_n) &\propto P(Y_k)P(y_1|Y_k)P(y_2|Y_k)P(y_3|Y_k)\cdots P(y_n|Y_k)\\ &\propto P(Y_k)\prod_{i=1}^{n}P(y_i|Y_k)\end{aligned}$$

接下来要用到 $\boldsymbol{X}$ 中包含的信息。我们通过 $\boldsymbol{X}$ 得到属于第 k 类对象的第 i 个特征满足分布密度函数 $f_k^{(i)}$，则上式化为：

$$P(Y_k)\prod_{i=1}^{n}P(y_i|Y_k)=\pi_k\prod_{i=1}^{n}f_k^{(i)}(y_i)$$

或者定义 f_k 是第 k 个分类的多元分布密度函数，这样上式可以化简为

$$P(Y_k)\prod_{i=1}^{n}P(y_i|Y_k)=\pi_k f_k(\overrightarrow{y})$$

那么这个式子是能够计算的了。使得该后验概率最大的 k 就是朴素 Bayes 方法得出的分类。

3.1.4 正态假设下的朴素 Bayes 分类 Naive Bayes Classification Under Normality

该模型需要一个较严格的假设：$f_k=N(\overrightarrow{\mu_k},\boldsymbol{\Sigma})$，注意到这里的方差矩阵和 k 无关。

则有结论：令下列函数最大的 k 等价于朴素 Bayes 得出的分类：

$$\sigma_k(\overrightarrow{y})=\overrightarrow{\mu_k}^T\boldsymbol{\Sigma}^{-1}\overrightarrow{y}-\frac{1}{2}\overrightarrow{\mu_k}^T\boldsymbol{\Sigma}^{-1}\overrightarrow{\mu_k}+\log\pi_k$$

3.2 线性判别分析 (Linear Discriminant Analysis, LDA)

3.3 二次判别分析 (Quadratic Discriminant Analysis, QDA)

第四章　差异检验模型

4.1　常用的假设检验模型

4.1.1　检验样本中的占比和总体中的占比是否有显著差异的模型

已知澳大利亚所有公民中患哮喘的比例是 7%，现抽取澳大利亚某一特定群体中的 200 人进行调查，发现 24 人患有哮喘，请问能不能认为这一特定群体的患病率与澳大利亚总体的患病率有显著差异？该样本患病率的置信区间是多少？（置信水平为 95%）

容易计算出这一特定人群中的患病率是 12%，我们把这一特定群体的患病率不显著偏离 7% 设定为原假设，把这一特定群体的患病率显著偏离 7% 设定为备择假设。现在按照之前给定的步骤计算 p-value 并将其和 5% 进行比较，看看这一事件发生的概率是不是不小于 5%。

此时使用之前提到的公式 $Z = \dfrac{\bar{X} - \mu}{SE}$ 计算，但是此时我们是使用 $\hat{p} - p_0$ 作为分子，$\hat{p}$ 代表的是被检验的概率大小（12%），p_0 则代表的是已知的、要进行对比的概率（7%）。这时的标准误差 $SE = \sqrt{\dfrac{p_0\,(1-p_0)}{n}}$，其中 n 是样本的大小，在这道题目中是 200。知道了以上参数后就可以计算出检验统计量 Z 为 2.77137。利用这个检验统计量符合正态分布的特性算出 $pvalue = P(|Z| \geq 2.77137) = 0.0056 < 0.05$，即原假设为真的概率是 0.0056，小于 5%，所以我们认为原假设在现实中不可能发生，从而拒绝原假设而接受备择假设，认为这一特定群体的患病率显著偏离 7%。

此时按照通用的置信区间计算公式：（**注意！计算置信区间时所用的 SE 不同！** $SE = \sqrt{\dfrac{\hat{p}\,(1-\hat{p})}{n}}$）

$$\big(\text{Estimate} - z_{\frac{\alpha}{2}} \times SE, \text{Estimate} + z_{\frac{\alpha}{2}} \times SE\big)$$

这里的估计值就是 $\hat{p}$，带入计算即可。

4.1.2　推测总体方差的模型

已知某容量为 n 的随机样本的标准差是 S^2，满足自由度为 $n-1$ 的卡方分布。选定显著性水平为 α，取得两个卡方分布的分位数 c_1 和 c_2 使得 $P(c_1 < S < c_2) = 1 - \alpha$，则它的一个置信水平为 $100(1-\alpha)\%$ 的置信区间为

$$\left(\frac{(n-1)S^2}{c_2}, \frac{(n-1)S^2}{c_1}\right)$$

但是这 c_1 和 c_2 的取法却有讲究。有无限多个 c_1 和 c_2 满足上述公式，要怎样选择才最好？有三种选择方式：

1. 选择对称的。即，选择 $P(S < c_1) = \dfrac{\alpha}{2}$，$P(S > c_2) = \dfrac{\alpha}{2}$，这是典型的双边检验。
2. 选择单边左的。即，选择 $c_1 = 0$，$P(S > c_2) = \alpha$，这是单边检验。
3. 选择单边右的。即，选择 $P(S < c_1) = \alpha$，$c_2 = +\infty$，这也是单边检验。

4.1.3 检验平均值是否和给定值有显著差异的模型

4.1.3.1 单样本 T 检验

猫血液中红细胞的含量服从平均值为 6.5 个单位的正态分布。现有 20 只猫组成的样本空间，对该样本中每只个体红细胞的含量进行测量，得到的平均值是 6.577 个单位，标准差是 0.8349983。现在想要知道，该样本的均值和总体的均值是否有显著差异？样本均值的置信区间是多少？（置信水平为 95%）

此时的检验统计量公式还是 $Z=\dfrac{\bar{X}-\mu}{SE}$，标准误差 SE 是 $\dfrac{s}{\sqrt{n}}$，其中 $s=\sqrt{\dfrac{1}{n-1}}\times\sum_{i=1}^{n}\left(X_i-\bar{X}\right)$，其中 $n-1$ 就是自由度（df）。

原假设是该样本的均值和总体均值无显著差异，备择假设是该样本的均值和总体均值有显著差异。检验统计量符合自由度为 $n-1$ 的 T 分布，计算出该检验统计量对应的 p-value 并和 5% 进行比较，即可决定是否拒绝原假设。

至于置信区间，按照通用的公式

$$\left(\text{Estimate}-z_{\frac{\alpha}{2}}\times SE,\text{Estimate}+z_{\frac{\alpha}{2}}\times SE\right)$$

直接带入计算即可。

4.1.3.2 标准差不同的独立双样本 T 检验 (非池化 (not pooled) 双样本 T 检验)

注意：标准差不同指的是两个样本的标准差相差超过 100%，即大的标准差大于小的标准差的两倍。

为了检验 HRT 激素替代疗法在治疗妇女的 HDL 胆固醇上的疗效，现有 60 名 75 岁以上的妇女作为实验对象被随机地分为两组，一组接受 HRT 激素替代疗法，另一组仅服用安慰剂。经过 9 个月的治疗，得到以下实验结果：（其中一名实验对象自然死亡）

	HRT 疗法组	安慰剂组
样本均值	8.1	2.4
样本标准差	10.5	4.3
样本容量	30	29

请问这两组对于 HDL 胆固醇的控制水平有没有显著差异?

此时认为原假设是 $H_0:\mu_1=\mu_2$，即两组控制水平没有显著差异，备择假设是 $H_a:\mu_1\neq\mu_2$。需要检验的参数是 $\bar{x}_1-\bar{x}_2$，并把其与零做比较（此时原假设也可以等价地叙述为 $H_0:\mu_1-\mu_2=0$），所以此时的检验统计量是 $Z=\dfrac{(\bar{x}_1-\bar{x}_2)-0}{SE}$。标准误差参见表格可以知道是 $SE=\sqrt{\dfrac{s_1^2}{n_1}+\dfrac{s_2^2}{n_2}}$，其中 s 代表对应的标准差，n 代表对应的样本空间。其遵循的是自由度为 $\min\{n_1,n_2\}-1$ 的 T 分布，即更少的样本数量减去一的 T 分布。计算可得到 p-value，和指定的显著性水平比较即可选择是否拒绝原假设。

选择自由度的时候，还有另外一种计算方式，可以令自由度等于

$$\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)}+\frac{s_2^4}{n_2^2(n_2-1)}}$$

在使用计算机进行假设检验时，可以用这个公式计算自由度。

4.1.3.3 标准差相同的独立双样本 T 检验（池化 (pooled) 双样本 T 检验）

注意：标准差相同指的是两个样本的标准差相差不超过 100%，即大的标准差不大于小的标准差的两倍。

现有 21 名志愿者被随机分为两个小组，其中一组接受钙补充剂，另外一组接受安慰剂，持续 12 周后分别测量他们的收缩压降低值，观察他们的收缩压降低有无显著差异，结果如下表所示。

接受类型	样本容量	平均降低收缩压	标准差
钙补充剂	10	5	8.743
安慰剂	11	-0.273	5.901

请问这两组之间的收缩压降低值有没有统计学意义上的显著差异？并计算置信区间。（置信水平 95%）

与上面的问题相似，这也是计算两组之间的某个值有没有显著区别，只是这一回两组的方差“认为相同”，标准误差变成了 $SE = s_P\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$，其中，$s_P = \sqrt{\frac{(n_1-1)\,s_1^2+(n_2-1)\,s_2^2}{n_1+n_2-2}}$，并且 $s_P \sim \chi^2_{n_1+n_2-2}$，自由度变为 n_1+n_2-2，其余步骤与上一个模型相同。

4.1.3.4 双样本 T 检验的非参版本——威尔科克逊-曼-惠特尼检验

给出用 A 和 B 两种不同疗法治疗的腺癌病人的存活时长（单位：月），希望知道两种疗法在存活时长上有无统计学意义上的显著差异。

Type A	7.02	5.60	6.59	20.14
Type B	16.81	21.67	13.4	29.11

由于存活时间不符合正态分布，所以在这里我们使用非参估计。这和双样本 T 检验事实上是等价的。

下面先介绍威尔科克逊检验。将所有测得的值从小到大排列：

组别	值	排序值
A	5.60	1
A	6.59	2
A	7.02	3
B	13.40	4
B	16.81	5
A	20.14	6
B	21.67	7
B	29.11	8

然后我们把 A 组或者 B 组的所有排序值加起来，得到 U。零假设是这两组数据没有显著区别（也就是说无论是哪一组的数据，每个数据比另外一组数据大或者小的概率都是二分之一），在这个假设下，我们可以计算出得到这个 U 值或者更极端的值的概率是多大（p-value），然后把这个概率和显著性水平进行比较。这和其他的模型一样，如果 p-value 比显著性水平更小，我们就说拒绝原假设；除此之外，我们保留原假设。

4.1.3.5 成对数据的检验

为了检验一种新型汽油是否对降低汽车耗油量有所帮助，某公司选择了 9 辆使用相同汽油类型的不同汽车，对每一辆汽车先后使用旧汽油和新型汽油并记录它们的每百公里耗油量，数据如下。

旧汽油每百公里耗油量	20.07	23.22	21.10	21.23	21.58	17.05	26.34	17.87	22.29
新型汽油每百公里耗油量	15.07	14.73	16.85	15.21	19.91	19.86	14.88	18.17	13.55
旧耗油量减新耗油量	5.63	8.49	4.25	6.02	1.67	-2.81	11.46	-0.30	8.74

所以新型汽油在统计学意义上有没有降低汽车的耗油量？（置信水平 95%）

其实我们只需要算出它们耗油量之差的均值和标准误差，然后把均值和零进行比较即可，遵循的分布是自由度为 $n-1$ 的 T 分布。概括地说，这和单样本 T 检验没有区别。

4.1.3.6 成对数据检验的非参版本

还是以上述数据作为例子。我们计算出新耗油量比旧耗油量少的数量，令零假设为新旧汽油在耗油量上没有区别（即二者耗油量比对方低的概率为二分之一），这样就得到一个二项分布 $B(n, 0.5)$。计算出在这个假设下新旧耗油量有别的概率大小并与显著性水平进行对比，然后选择是否拒绝原假设。

4.1.3.7 两个不同占比的检验

根据以下数据，分析美国男女大学生酗酒的比例有没有统计学上的显著差异?

性别组	样本容量	酗酒数量	酗酒比例 $\hat{p}$
男	5348	1392	0.26
女	8471	1748	0.206
总数	13819	3140	0.227

在这里我们说，如果男女大学生酗酒比例没有差别的话，那就是 $H_0 : \hat{p}_1 - \hat{p}_2 = 0$，备择假设则是比例有差别，即二者相减不为零。在这里我们想要检验的参数就是 $\hat{p}_1 - \hat{p}_2$，并且将它与零进行比较。此时的标准误差 $SE = \sqrt{\hat{p}\left(1-\hat{p}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$，再算出检验统计量即可。检验统计量遵循的分布是标准正态分布。

4.1.4 假设检验和置信区间总结表格

模型种类	参数	估计值	标准误差	对应分布
估计总体均值（总体标准差为 σ）	μ	$\bar{X}$	$\frac{\sigma}{\sqrt{n}}$	$N(0,1)$
估计总体均值（样本标准差为 s）	μ	$\bar{X}$	$\frac{s}{\sqrt{n}}$	$t(n-1)$
双样本 T 检验（标准差不相等）	$\mu_1-\mu_2$	$\bar{X}_1-\bar{X}_2$	$\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$	$t(min\{n_1,n_2\}-1)$
双样本 T 检验（标准差相等）	$\mu_1-\mu_2$	$\bar{X}_1-\bar{X}_2$	$\sqrt{\frac{(n_1-1)\,s_1^2+(n_2-1)\,s_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$	$t(n_1+n_2-2)$
成对数据	μ_D	$\bar{d}$	$\frac{s_D}{\sqrt{n}}$	$t(n-1)$
单占比置信区间	p	$\hat{p}=X/n$	$\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}$	$N(0,1)$
单占比差异检验	p	$\hat{p}=X/n$	$\sqrt{\frac{p_0\,(1-p_0)}{n}}$	$N(0,1)$
双占比置信区间	p_1-p_2	$\frac{X_1}{n_1}-\frac{X_2}{n_2}$	$\sqrt{\frac{\hat{p}_1\,(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2\,(1-\hat{p}_2)}{n_2}}$	$N(0,1)$
双占比差异检验	p_1-p_2	$\frac{X_1}{n_1}-\frac{X_2}{n_2}$	$\sqrt{\hat{p}\,(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$	$N(0,1)$

第五章　评估、诊断模型并对其进行改进

5.1　残差分析

1. 残差的期望值应该是零（线性）。即残差在图中应该大致均匀分布在零周围
2. 对于所有残差，从第一个残差往后取任意个残差，这些残差组中的方差应该随机分布在零周围（齐性）。即残差组的方差不随其容量的扩大而存在显著趋势。
3. 残差应该满足标准正态分布（正态性）。这一般通过 QQ 图来判断。

以上任意一个诊断标准不通过，都不能认为从这个模型中已经发掘出了足够的信息（即残差中还隐藏了部分信息）。

5.2　相关性系数

定义一个相关性系数 r^2:

$$r^2 = \frac{\sum_{i=1^n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2}{\sum_{i=1^n}(y_i - \bar{y})^2}$$

它的取值在正负一之间。它的绝对值越接近 1，说明线性相关性越强。一般绝对值超过 0.6 就可以说存在线性关系，超过 0.8 可以说有强烈的线性关系。有的时候我们会希望有一个调整过的 r_R^2，即

$$r_R^2 = 1 - \frac{n-1}{n-p-1}(1-r^2)$$

其中，p 是变量个数。

5.3　模型改进

如果以上诊断标准不通过怎么办？我们希望对模型进行改进以更好地符合上述诊断标准。

比如说，有的时候两个变量并不是线性关系，我们可以把原先的 x 替换成 $\ln x$ 或者是 e^x（对数、指数等都可以进行尝试，但记得要变形回去），然后再试。

第六章　方差分析 (ANOVA)

主要用来判断施加了不同影响的组之间的均值是否有明显差异。

6.1　单因素方差分析

现有一些羊被随机分为 5 组，其中一组为控制组，喂食普通的饲料，另外 4 组分别喂食 ABCD 四种新型饲料。过了一段时间后分别测量每组每只羊体重增长的大小，得到数据。现在想知道，新型饲料和普通饲料对于羊体重增长的效果究竟有没有区别?

零假设是 $H_0 : \mu_0 = \mu_1 = \mu_2 = \mu_3 = \mu_4$，备择假设是它们的增长量并不全部相等。

6.1.1　前提条件

为了进行单因素方差分析，我们必须要提前假设这些:

1. 这些组之间的数据互相独立，并且每组中的数据都大致符合正态分布（回忆 QQ 图）。
2. 这些组中最大和最小的方差之间不能相差超过一倍。

6.1.2　如何进行 F 检验

先看这个表格:

种类	方差和 (Sum of Squares, SS)	自由度 (DF)	均方差 (Mean Squares, MS)	F 检验量
不同组之间 (B)	$\sum_{i=1}^{k} n_i \left(\bar{Y}_{i\cdot\cdot} - \bar{Y}\right)^2$	k-1	SSB/DFB	MSB/MSW
组内 (W)	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{i\cdot\cdot}\right)^2$	n-k	SSW/DFW	
总共 (T)	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}\right)$	n-1	SST/DFT	

并且我们知道 F 检验量符合 F 分布 $F(k-1, n-k)$。这是一个有两个自由度的分布，第一个是 DFB，第二个是 DFW，**顺序不准颠倒**。

根据检验统计量符合的这个分布，我们进行通常的假设检验，计算 p-value，决定是否推翻原假设。

6.1.3　最小显著性差异法

如果原假设被拒绝，那么我们就知道并不是所有组中的均值都相等。但是如果我们只想知道某两组之间的均值是否有显著差异呢? 比如我们想知道 AB 两种饲料喂养的结果有没有显著差异，这时应该用什么办法呢?

这时我们用最小显著性差异法，检验统计量叫做 LSD，算法如下：

$$LSD = t_{\frac{\alpha}{2}}(n-k)\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

当两组均值之差（$|\bar{X}_i - \bar{X}_j|$）的绝对值大于 LSD 时，我们就说有显著差异。
有的时候为了避免假阳性，我们会进行一个 Bonferroni 修正，修正后的 LSD 看起来是这样的：

$$LSD = t_{\frac{\alpha}{2r}}(n-k)\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

其中，r 是排列组合数，$r = C_k^2$，k 是总的组数，这里是 5。

6.1.4 单因素方差分析在线性回归中的应用

6.1.5 双因素方差分析

第七章　附录 Appendix

7.1　证明

7.1.1　随机向量方差的性质

容易证明常数项 $\overrightarrow{b}$ 不影响方差。设 A_{ij} 是矩阵 $\boldsymbol{A}$ 第 i 行第 j 列的元素，接下来证明 $\mathrm{Var}\left(()\,\boldsymbol{A}\overrightarrow{X}\right)=\boldsymbol{A\Sigma A}^T$：

$$
\begin{aligned}
\mathrm{Var}\left(\boldsymbol{A}\overrightarrow{X}\right) &= \mathrm{Var}\left(\left(\sum_{i=1}^{n}A_{1i}X_i,\sum_{i=1}^{n}A_{2i}X_i,\sum_{i=1}^{n}A_{2i}X_i,\ldots,\sum_{i=1}^{n}A_{mi}X_i\right)\right)\\
&=\begin{pmatrix}
\mathrm{Cov}\left(\sum_{i=1}^{n}A_{1i}X_i,\sum_{i=1}^{n}A_{1i}X_i\right) & \mathrm{Cov}\left(\sum_{i=1}^{n}A_{1i}X_i,\sum_{i=1}^{n}A_{2i}X_i\right) & \cdots & \mathrm{Cov}\left(\sum_{i=1}^{n}A_{1i}X_i,\sum_{i=1}^{n}A_{mi}X_i\right)\\
\mathrm{Cov}\left(\sum_{i=1}^{n}A_{2i}X_i,\sum_{i=1}^{n}A_{1i}X_i\right) & \mathrm{Cov}\left(\sum_{i=1}^{n}A_{2i}X_i,\sum_{i=1}^{n}A_{2i}X_i\right) & \cdots & \mathrm{Cov}\left(\sum_{i=1}^{n}A_{2i}X_i,\sum_{i=1}^{n}A_{mi}X_i\right)\\
\mathrm{Cov}\left(\sum_{i=1}^{n}A_{3i}X_i,\sum_{i=1}^{n}A_{1i}X_i\right) & \mathrm{Cov}\left(\sum_{i=1}^{n}A_{3i}X_i,\sum_{i=1}^{n}A_{2i}X_i\right) & \cdots & \mathrm{Cov}\left(\sum_{i=1}^{n}A_{3i}X_i,\sum_{i=1}^{n}A_{mi}X_i\right)\\
\vdots & \vdots & \ddots & \vdots\\
\mathrm{Cov}\left(\sum_{i=1}^{n}A_{mi}X_i,\sum_{i=1}^{n}A_{1i}X_i\right) & \mathrm{Cov}\left(\sum_{i=1}^{n}A_{mi}X_i,\sum_{i=1}^{n}A_{2i}X_i\right) & \cdots & \mathrm{Cov}\left(\sum_{i=1}^{n}A_{mi}X_i,\sum_{i=1}^{n}A_{mi}X_i\right)
\end{pmatrix}\\
&=\begin{pmatrix}
\sum_{i=1}^{n}\sum_{j=1}^{n}A_{1i}A_{1j}\mathrm{Cov}\left(X_i,X_j\right) & \sum_{i=1}^{n}\sum_{j=1}^{n}A_{1i}A_{2j}\mathrm{Cov}\left(X_i,X_j\right) & \cdots & \sum_{i=1}^{n}\sum_{j=1}^{n}A_{1i}A_{mj}\mathrm{Cov}\left(X_i,X_j\right)\\
\sum_{i=1}^{n}\sum_{j=1}^{n}A_{2i}A_{1j}\mathrm{Cov}\left(X_i,X_j\right) & \sum_{i=1}^{n}\sum_{j=1}^{n}A_{2i}A_{2j}\mathrm{Cov}\left(X_i,X_j\right) & \cdots & \sum_{i=1}^{n}\sum_{j=1}^{n}A_{2i}A_{mj}\mathrm{Cov}\left(X_i,X_j\right)\\
\sum_{i=1}^{n}\sum_{j=1}^{n}A_{3i}A_{1j}\mathrm{Cov}\left(X_i,X_j\right) & \sum_{i=1}^{n}\sum_{j=1}^{n}A_{3i}A_{2j}\mathrm{Cov}\left(X_i,X_j\right) & \cdots & \sum_{i=1}^{n}\sum_{j=1}^{n}A_{3i}A_{mj}\mathrm{Cov}\left(X_i,X_j\right)\\
\vdots & \vdots & \ddots & \vdots\\
\sum_{i=1}^{n}\sum_{j=1}^{n}A_{mi}A_{1j}\mathrm{Cov}\left(X_i,X_j\right) & \sum_{i=1}^{n}\sum_{j=1}^{n}A_{mi}A_{2j}\mathrm{Cov}\left(X_i,X_j\right) & \cdots & \sum_{i=1}^{n}\sum_{j=1}^{n}A_{mi}A_{mj}\mathrm{Cov}\left(X_i,X_j\right)
\end{pmatrix}\\
&=\boldsymbol{A\Sigma A}^T
\end{aligned}
$$

证毕。

7.1.2 唯一最小值的取得

通过公式可以获知：

$$\begin{aligned} q(\nu) &= \sum_{i=1}^{n}(u_i - \nu) \\ &= \sum_{i=1}^{n}(u_i^2 - 2u_i\nu + \nu^2) \\ &= \sum_{i=1}^{n} u_i^2 - 2\sum_{i=1}^{n} u_i\nu + \sum_{i=1}^{n}\nu^2 \end{aligned}$$

然后对其求一阶导

$$\frac{\mathrm{d}q}{\mathrm{d}\nu} = -2\sum_{i=1}^{n} u_i + 2\sum_{i=1}^{n}\nu$$

则当 $\nu = \bar{u}$ 时，导数等于零。接下来求二阶导，证明导数等于零时，函数取最小值。

$$\frac{\mathrm{d}^2 q}{\mathrm{d}\nu^2} = 2n > 0$$

所以函数在 $\nu = \bar{u}$ 时取得最小值，证毕。

7.1.3 线性回归系数矩阵的计算

我们的目的是令 $\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$ 取到最小值。将它展开，得到：

$$\begin{aligned} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 &= (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= (\boldsymbol{Y}^T - \boldsymbol{\beta}^T\boldsymbol{X}^T)(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \quad \text{（注意到这里每一项都是方阵）} \\ &= \boldsymbol{Y}^T\boldsymbol{Y} - 2\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \end{aligned}$$

接下来对 $\boldsymbol{X}$ 进行特殊的 QR 分解，即令 $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$，其中 $\boldsymbol{Q}$ 是尺寸为 $n \times (p+1)$ 满足 $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}$ 的矩阵，$\boldsymbol{R}$ 是 $(p+1) \times (p+1)$ 的上三角阵。完成此操作要求 $\boldsymbol{X}$ 的列向量满秩。将分解后的矩阵代入上式，得到：

$$\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = \boldsymbol{Y}^T\boldsymbol{Y} - 2\boldsymbol{Y}^T\boldsymbol{Q}\boldsymbol{R}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{R}^T\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{R}\boldsymbol{\beta}$$

在这个式子的右边同时加上和减去 $\boldsymbol{Y}^T\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{Y}$，使我们可以将右侧写成两项之和，其中只有一项包含 $\boldsymbol{\beta}$：

$$\begin{aligned} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 &= \boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{Y} + (\boldsymbol{Y}^T\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{Y} - 2\boldsymbol{Y}^T\boldsymbol{Q}\boldsymbol{R}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{R}^T\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{R}\boldsymbol{\beta}) \\ &= \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{Q}\boldsymbol{Q}^T)\boldsymbol{Y} + (\boldsymbol{Q}^T\boldsymbol{Y} - \boldsymbol{R}\boldsymbol{\beta})^T(\boldsymbol{Q}^T\boldsymbol{Y} - \boldsymbol{R}\boldsymbol{\beta}) \end{aligned}$$

要令上式取到最小值，需要使 $\boldsymbol{Q}^T\boldsymbol{Y} - \boldsymbol{R}\boldsymbol{\beta} = 0$，即 $\boldsymbol{Q}^T\boldsymbol{Y} = \boldsymbol{R}\boldsymbol{\beta}$. 因为 $\boldsymbol{X}$ 的列满秩，所以 $\boldsymbol{R}$ 一定可逆，得到 $\hat{\boldsymbol{\beta}} = \boldsymbol{R}^{-1}\boldsymbol{Q}^T\boldsymbol{Y}$.

然后我们证明这个结论和所证明的结论等价：

$$\begin{aligned} \boldsymbol{X}^T &= \boldsymbol{R}^T\boldsymbol{Q}^T \\ \boldsymbol{X}^T\boldsymbol{X} &= \boldsymbol{R}^T\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{R} = \boldsymbol{R}^T\boldsymbol{R} \\ (\boldsymbol{X}^T\boldsymbol{X})^{-1} &= \boldsymbol{R}^{-1}(\boldsymbol{R}^T)^{-1} \\ (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T &= \boldsymbol{R}^{-1}(\boldsymbol{R}^T)^{-1}\boldsymbol{R}^T\boldsymbol{Q}^T = \boldsymbol{R}^{-1}\boldsymbol{Q}^T \end{aligned}$$

所以 $\hat{\boldsymbol{\beta}} = \boldsymbol{R}^{-1}\boldsymbol{Q}^T\boldsymbol{Y} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$，证毕。

7.1.4 最小二乘方差的估计

为了证明这个结论，首先做准备工作：

$$\begin{aligned}\mathrm{tr}\,(\boldsymbol{H}) &= \mathrm{tr}\left(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right)\\ &= \mathrm{tr}\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\right)\\ &= \mathrm{tr}\,(\boldsymbol{I}_{p+1})\\ &= p+1\end{aligned}$$

所以得到 $\sum_{i=1}^{n} h_{ii} = p+1$. 得到这个结论后，又有：

$$\begin{aligned}\mathrm{Var}\,(\boldsymbol{\epsilon}) &= \mathrm{Var}\,((\boldsymbol{I}-\boldsymbol{H})\overrightarrow{y})\\ &= \mathrm{Cov}\,((\boldsymbol{I}-\boldsymbol{H})\overrightarrow{y},(\boldsymbol{I}-\boldsymbol{H})\overrightarrow{y})\\ &= (\boldsymbol{I}-\boldsymbol{H})\mathrm{Cov}\,(\overrightarrow{y},\overrightarrow{y})\,(\boldsymbol{I}-\boldsymbol{H})^T\\ &= (\boldsymbol{I}-\boldsymbol{H})\sigma^2\boldsymbol{I}(\boldsymbol{I}-\boldsymbol{H})^T\\ &= \sigma^2(\boldsymbol{I}-\boldsymbol{H})\end{aligned}$$

所以得到：

$$\begin{aligned}\mathrm{E}\left[\frac{1}{n-p-1}\sum_{i=1}^{n}\epsilon_i^2\right] &= \frac{1}{n-p-1}\mathrm{E}\left[\sum_{i=1}^{n}\epsilon_i^2\right]\\ &= \frac{1}{n-p-1}\left[\mathrm{Var}\,(\epsilon_i)+(\mathrm{E}\,[\epsilon_i])^2\right]\\ &= \frac{1}{n-p-1}\sum_{i=1}^{n}\mathrm{Var}\,(\epsilon_i)\end{aligned}$$

已知 $\mathrm{Var}\,(\boldsymbol{\epsilon}) = \sigma^2(\boldsymbol{I}-\boldsymbol{H})$，所以 $\epsilon_i = (1-h_{ii})\sigma^2$，所以上式变为：

$$\begin{aligned}\mathrm{E}\left[\frac{1}{n-p-1}\sum_{i=1}^{n}\epsilon_i^2\right] &= \frac{1}{n-p-1}\sum_{i=1}^{n}\sigma^2(1-h_{ii})\\ &= \frac{1}{n-p-1}\left(n-\sum_{i=1}^{n}h_{ii}\right)\sigma^2\\ &= \frac{1}{n-p-1}(n-(p+1))\sigma^2\\ &= \sigma^2\end{aligned}$$

所以该估计量是无偏的。接下来证明它符合卡方分布。因为：

$$\frac{(n-p-1)s_e^2}{\sigma^2} = \sum_{i=1}^{n}\frac{\epsilon_i^2}{\sigma^2}$$

而且根据假设，有：

$$\epsilon_i \sim N(0,\sigma^2)$$

所以根据卡方分布的定义，有 $\dfrac{(n-p-1)s_e^2}{\sigma^2} \sim \chi^2(n-p-1)$，证毕。

7.1.5 最小二乘系数矩阵的统计性质

期望:

$$\mathrm{E}\left[\hat{\boldsymbol{\beta}}\right] = \mathrm{E}\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}\right] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathrm{E}\left[\boldsymbol{Y}\right] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

方差:

$$\begin{aligned}\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right) &= \mathrm{Var}\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}\right)\\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathrm{Var}\left(\boldsymbol{Y}\right)\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\sigma^2\boldsymbol{I}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\end{aligned}$$

接下来证明分布。首先我们根据假设，有:

$$\boldsymbol{\epsilon} \sim N(\overrightarrow{0}, \sigma^2\boldsymbol{I})$$

因为 $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$，所以有:

$$\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$$

而 $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$，所以 $\hat{\boldsymbol{\beta}}$ 是 $\boldsymbol{Y}$ 的线性组合，是一个符合正态分布的随机变量的线性组合。又因为 $\hat{\boldsymbol{\beta}}$ 的期望和方差已知，所以有:

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$$

证毕。

7.1.6 Gauss-Markov Theorem

由于证明的重点在于 MVLUE，所以我们不妨假设有另一个线性无偏估计量，即 $\Lambda(\boldsymbol{y})$，他是无偏的，即:

$$\mathrm{E}\left[\Lambda(\boldsymbol{y})\right] = \Lambda(\boldsymbol{X}\boldsymbol{\beta}) = \lambda(\boldsymbol{\beta})$$

所以有 $\Lambda(\boldsymbol{X}) = \lambda$，即这两个线性变换是相同的。我们约定在该证明中将 Λ 以及 λ 视作两个矩阵，因为矩阵本质上就是线性变换，所以刚才的式子可转写为 $\Lambda\boldsymbol{X} = \lambda$，这样看起来就和谐多了。

然后有:

$$\mathrm{Var}\left(\Lambda\boldsymbol{y}\right) = \Lambda\mathrm{Var}\left(\boldsymbol{y}\right)\Lambda^T = \sigma^2\Lambda\Lambda^T$$

并且:

$$\begin{aligned}\mathrm{Var}\left(\lambda\hat{\boldsymbol{\beta}}\right) &= \mathrm{Var}\left(\lambda(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}\right)\\ &= \lambda(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathrm{Var}\left(\boldsymbol{y}\right)\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\lambda^T\\ &= \sigma^2\lambda(\boldsymbol{X}^T\boldsymbol{X})^{-1}\lambda^T\\ &= \sigma^2\Lambda\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\Lambda^T\end{aligned}$$

于是:

$$\begin{aligned}\mathrm{Var}\left(\Lambda\boldsymbol{y}\right) - \mathrm{Var}\left(\lambda\hat{\boldsymbol{\beta}}\right) &= \sigma^2\Lambda(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)\Lambda^T\\ &= \sigma^2\Lambda(\boldsymbol{I} - \boldsymbol{H})\Lambda^T\\ &\geq 0\end{aligned}$$

其中，$(\boldsymbol{I} - \boldsymbol{H})$ 是投影阵，所以它是非负定的，证毕。

7.1.7 残差方差的另外一种计算方式

$$
\begin{aligned}
S_e^2 &= \frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \\
&= \frac{1}{n-2}\sum_{i=1}^{n}[(y_i-(\hat{\beta}_0+\hat{\beta}_1 x_i))^2] \\
&= \frac{1}{n-2}\sum_{i=1}^{n}(y_i-\bar{y}+\hat{\beta}_1\bar{x}-\hat{\beta}_1 x_i)^2 \\
&= \frac{1}{n-2}\sum_{i=1}^{n}[(y_i-\hat{y})-(x_i-\bar{x})\hat{\beta}_1]^2 \\
&= \frac{1}{n-2}\sum_{i=1}^{n}[(y_i-\bar{y})^2-2\hat{\beta}_1(y_i-\hat{y})(x_i-\bar{x})+(x_i-\bar{x})^2\hat{\beta}_1^2] \\
&= \frac{1}{n-2}(S_{yy}-2\hat{\beta}_1 S_{xy}+S_{xx}\hat{\beta}_1^2)
\end{aligned}
$$

因为 $\hat{\beta}_1=\frac{S_{xy}}{S_{xx}}$，所以 $\hat{\beta}_1 S_{xy}=\frac{S_{xy}^2}{S_{xx}}=\hat{\beta}_1^2 S_{xx}$，代入得到

$$
\begin{aligned}
S_e^2 &= \frac{1}{n-2}(S_{yy}-2\hat{\beta}_1^2 S_{xx}+\hat{\beta}_1^2 S_{xx}) \\
&= \frac{1}{n-2}(S_{yy}-\hat{\beta}_1^2 S_{xx})
\end{aligned}
$$

证毕。

7.1.8 最小二乘法的模型系数分布证明

首先我们要对最小二乘法（2.2）中要用到的几个参数进行变形操作，以便于后续证明。

$$
\begin{aligned}
S_{xy} &= \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \\
&= \sum_{i=1}^{n}[(x_i-\bar{x})y_i-(x_i-\bar{x})\bar{y}] \\
&= \sum_{i=1}^{n}(x_i-\bar{x})y_i-\bar{y}\sum_{i=1}^{n}(x_i-\bar{x}) \\
&= \sum_{i=1}^{n}(x_i-\bar{x})y_i
\end{aligned}
$$

类似地，我们有

$$
S_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})x_i
$$

利用化简过后的 S_{xx} 与 S_{xy}，我们再计算 $\hat{\beta}_0$ 以及 $\hat{\beta}_1$：

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} \\
&= \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) y_i \\
&\text{令 } a_i = \frac{(x_i - \bar{x})}{S_{xx}}\text{，则原式} \\
&= \sum_{i=1}^{n} a_i y_i \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\
&= \frac{1}{n} \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a_i y_i \bar{x} \\
&= \sum_{i=1}^{n} (\frac{1}{n} - a_i \bar{x}) y_i \\
&\text{令 } b_i = \frac{1}{n} - a_i \bar{x}\text{，则原式} \\
&= \sum_{i=1}^{n} b_i y_i
\end{aligned}
$$

然后我们先证明 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的期望。注意到下面的证明使用了上述定义的一些变量。

$$
\begin{aligned}
\mathrm{E}\left[\hat{\beta}_1\right] &= \sum_{i=1}^{n} a_i \mathrm{E}\left[y_i\right] \\
&= \sum_{i=1}^{n} \frac{x_i - \bar{x}}{S_{xx}} (\beta_0 + \beta_1 x_i) \\
&= \frac{1}{S_{xx}} \left[\beta_0 \sum_{i=1}^{n} (x_i - \bar{x}) + \beta_1 \sum_{i=1}^{n} (x_i - \bar{x}) x_i \right] \\
&\text{因为 } \sum_{i=1}^{n} (x_i - \bar{x}) = 0\text{，并且 } \sum_{i=1}^{n} (x_i - \bar{x}) x_i = S_{xx}\text{，所以原式} \\
&= \frac{1}{S_{xx}} \beta_1 S_{xx} \\
&= \beta_1 \\
\mathrm{E}\left[\hat{\beta}_0\right] &= \sum_{i=1}^{n} \mathrm{E}\left[y_i\right] \\
&= \sum_{i=1}^{n} \left[\left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{S_{xx}} \right) (\beta_0 + \beta_1 x_i) \right] \\
&= \frac{1}{n} \sum_{i=1}^{n} \beta_0 - \frac{\beta_0 \bar{x}}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) + \frac{\beta_1}{n} \sum_{i=1}^{n} x_i - \frac{\beta_1 \bar{x}}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) x_i \\
&= \frac{1}{n} \sum_{i=1}^{n} \beta_0 - \frac{\beta_0 \bar{x}}{S_{xx}} \times 0 + \frac{\beta_1}{n} n \bar{x} - \frac{\beta_1 \bar{x}}{S_{xx}} S_{xx} \\
&= \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} \\
&= \beta_0
\end{aligned}
$$

接下来我们证明它们的方差，其中仍然使用到了上述定义的两个变量 a_i 和 b_i：

$$
\begin{aligned}
\operatorname{Var}\left(\hat{\beta}_1\right) &= \sum_{i=1}^{n} a_i^2(y_i) \\
&= \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S_{xx}}\right)^2 \sigma^2 \\
&= \frac{\sigma^2}{S_{xx}}^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= \frac{\sigma^2}{S_{xx}}^2 S_{xx} \\
&= \frac{\sigma^2}{S_{xx}} \\
\operatorname{Var}\left(\hat{\beta}_0\right) &= \sum_{i=1}^{n} b_i^2 \operatorname{Var}(y_i) \\
&= \sum_{i=1}^{n} \left[\left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{S_{xx}}\right)^2 \sigma^2\right] \\
&= \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n^2} - \frac{2\bar{x}(x_i - \bar{x})}{nS_{xx}} + \frac{\bar{x}^2 (x_i - \bar{x})^2}{S_{xx}^2}\right) \\
&= \sigma^2 \left(\frac{1}{n} - \frac{2\bar{x}}{nS_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) + \frac{\bar{x}^2}{S_{xx}^2} \sum_{i=1}^{n} (x_i - \bar{x})^2\right) \\
&= \sigma^2 \left(\frac{1}{n} - \frac{2\bar{x}}{nS_{xx}} \times 0 + \frac{\bar{x}^2}{S_{xx}^2} S_{xx}\right) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)
\end{aligned}
$$

7.1.9 最小二乘法系数的协方差证明

$$
\begin{aligned}
\operatorname{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) &= \operatorname{Cov}\left(\sum_{i=1}^{n} b_i y_i, \sum_{j=1}^{n} a_j y_j\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} b_i a_j \operatorname{Cov}\left(y_i, y_j\right) \\
&= \sum_{i=1}^{n}\left[b_i a_i \operatorname{Cov}\left(y_i, y_i\right) + \sum_{\substack{i=1\\ j=1\\ i\neq j}}^{n} b_i a_j \operatorname{Cov}\left(y_i, y_j\right)\right] \\
&= \sum_{i=1}^{n}\left[b_i a_i \operatorname{Var}\left(y_i\right) + \sum_{\substack{i=1\\ j=1\\ i\neq j}}^{n} b_i a_j \times 0\right] \text{（已知 } y_i \text{ 之间相互独立，所以协方差等于零）} \\
&= \sum_{i=1}^{n} a_i b_i \operatorname{Var}\left(y_i\right) \\
&= \sigma^2 \sum_{i=1}^{n} \frac{x_i - \bar{x}}{S_{xx}}\left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{S_{xx}}\right) \\
&= \frac{\sigma^2}{S_{xx}}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) - \frac{\bar{x}}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})^2\right] \\
&= \frac{\sigma^2}{S_{xx}}\left[\frac{1}{n} \times 0 - \frac{\bar{x}}{S_{xx}} S_{xx}\right] \\
&= \frac{\sigma^2}{S_{xx}} \times -\bar{x} \\
&= -\frac{\sigma \bar{x}^2}{S_{xx}}
\end{aligned}
$$

7.1.10 最小二乘法系数分布的证明 1

（此证明暂时不要求掌握）

7.1.11 最小二乘法系数分布的证明 2

（此证明暂时不要求掌握）

7.1.12 预测值分布的证明 1

7.1.13 预测值分布的证明 2

7.1.14 预测值分布的证明 3

7.1.15 预测区间和置信区间的不同

7.1.16 残差满足的性质

7.1.17 残差的标准化与学生化

（此证明不要求掌握）

7.1.18 列向量必须线性无关

7.1.19 列线性无关的矩阵性质 1

7.1.20 多元最小二乘法系数的计算

7.1.21 多元最小二乘法残差的计算