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

Leo_Maxwell

March 2023

Good Explanation>Symbolic Proof.

目录

序言 Introduction

本文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。

1 常用的假设检验模型

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！** $SE = \sqrt{\dfrac{\hat{p}\,(1-\hat{p})}{n}}$）

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

这里的估计值就是 $\hat{p}$，带入计算即可。

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。

1.3 检验平均值是否和给定值有显著差异的模型

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。

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。

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。

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。

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。

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。

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。

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	$\frac{X_1}{n_1}-\frac{X_2}{n_2}$	$\sqrt{\hat{p}\,(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$	$N(0,1)$

2 线性模型 (Linear Model, LM)

2.1 基本概念

有两个变量，这两个变量之间**存在联系，但不是函数关系**，但是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相关性（即拟合）。例如，一般来说体重越重的人身高也越高，但是究竟大概是什么样的关系呢？我们希望用一个函数来描述这种关系。而一元线性回归就是指对于两个变量之间线性的关系进行拟合（其中一个变量对另外一个变量造成影响）。

现在有许多组数据，(x_1,y_1), (x_2,y_2) ,..., (x_n,y_n)，在简单的一元线性回归中，我们希望找到一个这样的函数关系:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

其中，Y_i 是观测数据中真正的值，β_0 是截距，β_1 是斜率，最为重要的是 ε_i，它是残差，也是这个公式中的“随机组成部分”，即拟合值与真实值之间的差距。如果把残差去掉，剩下的部分就是一个简单的一次函数。而且

我们希望，在这样的模型下，对于残差而言，有 $\varepsilon_i \sim N(0,\sigma^2)$，并且相互独立。这样的模型叫做简单线性回归模型 (Simple Linear Regression Model, SLR Model)。在后面我们会知道无论是什么关系都可以转换成线性关系，从而构建线性模型。

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

使得下列二元函数最小的 β_0 和 β_1，就是用最小二乘法得出的理想的 β_0 和 β_1，我们记为 $\hat{\beta}_0$ 和 $\hat{\beta}_1$。

$$Q(\beta_0,\beta_1)=\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1 x_i))^2$$

所以其得名最小二乘法。因为若 $u_1,u_2,u_3,\ldots,u_n$ 是一列实数，则函数 $q(\gamma)=\sum_{i=1}^{n}(u_i-\gamma)^2$ **仅在** $\gamma=\bar{u}$ 时取得最小值（证明请见附录5.1.1），所以我们能通过以下方式计算得到 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的值，证明请见附录5.1.2。

2.2.1 系数计算和性质

对于 $\hat{\beta}_1$ 和 $\hat{\beta}_0$，我们有

$$\hat{\beta}_1=\frac{S_{xy}}{S_{xx}}$$
$$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}$$

其中，S_{xy} 被称作 cross product sum of squares，S_{xx} 被称作 sum of squares due to x，计算方式如下：

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

$$S_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2$$

现在我们回到概率统计的框架中来。因为线性回归是对许多组成对数据 (x_i,y_i) 进行操作，而这些成对数据中的 y_i 一般都是在相同 x_i 的条件下多次重复实验得到的，或者是从一个未知分布中随机抽取得到的值（例如从一群身高相同的人中抽取一个人的体重），所以线性回归的对象一般都是一列互相独立的随机变量 $Y_1,Y_2,\ldots,Y_n$，我们假设它们满足 $\mathrm{E}[Y_i]=\beta_0+\beta_1 x_i$ 和 $\mathrm{Var}(Y_i)=\sigma^2$，其中，$x_i$ 就是上述许多组成对数据中的 x_i。于是，$\hat{\beta}_0$ 和 $\hat{\beta}_1$ 就从两个数字变成两个随机变量。作为随机变量，它们必然是拥有随机变量的属性的，在这里给出：（证明请见附录5.1.5）

	期望	方差
$\hat{\beta}_0$	β_0	$\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)$
$\hat{\beta}_1$	β_1	$\frac{\sigma^2}{S_{xx}}$

特别地，$\mathrm{Cov}(\hat{\beta}_0,\hat{\beta}_1)=-\frac{\bar{x}\sigma^2}{S_{xx}}$，证明请见5.1.6。

如果我们的数据能够满足 $Y_i\sim N(\beta_0+\beta_1 x_i,\sigma^2)$，且每个 Y_i 之间相互独立，那么还有性质：（证明请见附录5.1.7）

$$\hat{\beta}_0\sim N\left(\beta_0,\sigma^2\left(\frac{1}{n}+\frac{\bar{x}}{S_{xx}}\right)\right)$$
$$\hat{\beta}_1\sim N\left(\beta_1,\frac{\sigma^2}{S_{xx}}\right)$$
$$\frac{(n-2)S_e^2}{\sigma^2}\sim\chi^2_{n-2}$$

2.2.2 残差的性质

在定义中我们提到，对我们的线性回归模型而言，残差符合正态分布 $N(0,\sigma^2)$。所以在这里残差也是随机变量，只要是随机变量，我们就研究它的性质。该分布中只有一个参数是我们未知的，所以我们接下来推测这个参数 σ^2。

我们仍然用样本的方差 S_e^2 来估计总体方差 σ^2，如同我们之前做过的那样：（说明请见附录5.1.3）

$$S_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

其中，$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$，且

$$\mathrm{E}[S_e^2] = \sigma^2$$

并且还有：（证明请见附录5.1.4）

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

残差的另外一个重要应用是用来检测建立模型时的假设是否成立。上述建立模型的通常假设是 $Y_1, Y_2, \ldots, Y_n$ 满足模型 $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$，并且 ε_i 之间相互独立，还有 $\varepsilon_i \sim N(0,\sigma^2)$。现在我们用 $\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_n$ 来表示残差，它们的定义是 $\hat{e}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$。

现在我们知道残差满足以下的性质：（证明请见附录5.1.13）

- $\sum_{i=1}^{n} \hat{e}_i = 0$
- $\sum_{i=1}^{n} \hat{e}_i x_i = 0$
- $\mathrm{E}[\hat{e}_i] = 0$
- $\mathrm{Var}(\hat{e}_i) = \sigma^2\left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}\right)$

并且我们把以下公式称作是标准化残差 (Standardized Residuals)：

$$\tilde{e}_i = \frac{y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)}{\sigma\sqrt{1 - \frac{1}{n} - \frac{(x_i-\bar{x})^2}{S_{xx}}}}$$

而将以下公式称作是学生化残差 (Studentized Residuals)：

$$e_i^* = \frac{y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)}{S_e\sqrt{1 - \frac{1}{n} - \frac{(x_i-\bar{x})^2}{S_{xx}}}}$$

严格地说，这个学生化残差称作是学生化内残差 (Internally Studentized Residuals)，并不符合 T 分布。我们还可以利用学生化外残差 (Externally Studentized Residuals) 来解决这个问题。在计算学生化外残差时，上述公式不变，但是使用的线性模型必须去除第 i 个数据。有关于标准化和学生化残差的详细证明，请见附录5.1.14。

2.2.3 系数符合的分布

知道了残差的方差以及系数的计算方式后，我们有：（证明请见5.1.8）

$$\frac{\hat{\beta}_0 - \beta_0}{S_e\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim t_{n-2}$$

$$\frac{\hat{\beta}_1 - \beta_1}{S_e/\sqrt{S_{xx}}} \sim t_{n-2}$$
$$\frac{(n-2)S_e^2}{\sigma^2} \sim \chi^2_{n-2}$$

其实上述服从 T 分布的就是检验统计量 Z，形式和之前假设检验是一样的。分子是估计值减去参数，分母是标准误差。置信区间的算法也和之前的一模一样，此处不再赘述。

2.3 预测与预测区间

得到了关于许多个已知期望和方差，但符合未知分布的 Y_i 的观察值 y_i 的拟合线条以后，我们会希望预测对于一个给定 x_0，对应的 Y_0 有什么性质。给定一个 x_0，则 $Y_0 = \beta_0 + \beta_1 x_0$，并且满足：(证明请见附录5.1.9)

- $\mathrm{E}[\hat{\beta}_0 + \hat{\beta}_1 x_0] = \beta_0 + \beta_1 x_0$
- $\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$

特别地，如果每个 Y_i 都满足已知正态分布 $N(\beta_0 + \beta_1 x_i, \sigma^2)$，则有：（证明请见附录5.1.10）

$$\hat{\beta}_0 + \hat{\beta}_1 \sim N\left(\beta_0 + \beta_1 x_0, \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) \right)$$

并且在满足 Y_i 正态性的前提下，我们还有：（证明请见附录5.1.11）

$$\frac{(\hat{\beta}_0 + \hat{\beta}_1 x_0) - (\beta_0 + \beta_1 x_0)}{S_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}$$

我们在之前的学习中已经知道有关于置信区间的知识。对于类似于 $\beta_0 + \beta_1 x_0$ 这样已知分布的随机变量，我们同样可以计算出这个随机变量在某个置信水平下覆盖的置信区间。它是

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{\frac{\alpha}{2}, n-2} S_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

其中，x_0 是回归曲线上的某个值，S_e 是残差的标准差。$t_{\frac{\alpha}{2}, n-2}$ 是自由度为 $n-2$ 的 T 分布上的分位数。

这个区间叫做是 $\beta_0 + \beta_1 x_0$ 的**置信区间** (Confidence Interval, CI)。现在我们要计算 Y_0 的**预测区间** (Prediction Interval, PI) 的话，公式有一些小小的变动：

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{\frac{\alpha}{2}, n-2} S_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

注意到预测区间和置信区间有本质上的不同，类似于 $\beta_0 + \beta_1 x_0$ 和 Y_0 之间也有本质的不同一样。这里的置信区间是基于**估计**的方差而得出的，但是预测区间是根据**预测**的方差得出的。二者很大程度上是不同的，具体证明可见附录5.1.12。

2.4 多元线性回归模型 (Multiple Linear Regression, MLR)

和一元线性回归模型类似，只是 MLR 中是几个变量一同决定单个变量。这些变量之间还可能组成乘积等形式（即相互作用）。最简单的一个 MLR 是

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_1 X_{i2} + \cdots + \beta_r X_{ir} + \varepsilon_i$$

2.5 矩阵表示

因为有多个观察数据，所以有多个拟合结果。由线性代数知识易知可以由矩阵表示多元线性方程组。
例如，拥有 r 个变元的简单线性回归模型，可以用这样的矩阵表示：

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1r} \\ 1 & x_{21} & x_{22} & x_{23} & \cdots & x_{2r} \\ 1 & x_{31} & x_{32} & x_{33} & \cdots & x_{3r} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nr} \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_r \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

化简为

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

2.6 性质

$\boldsymbol{X}$ 中的列向量必须线性无关。若该性质不满足，则无法确定唯一的 $\boldsymbol{\beta}$。该性质的证明请见附录5.1.15。
若矩阵 $\boldsymbol{X}_{n\times p}$ 中的列向量线性无关，则 $\boldsymbol{X}^T\boldsymbol{X}$ 矩阵可逆。证明请见附录5.1.16。

2.7 最小二乘法

和之前的定义类似。我们希望找到 $r+1$ 个值 $\beta_0, \beta_1, \beta_2, \ldots, \beta_r$ 使得这个函数最小：

$$Q(\beta_0, \beta_1, \beta_2, \ldots . \beta_r) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_r x_{ir} \varepsilon_i))^2$$

并且我们能计算，$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$，证明请见附录5.1.17。
残差的方差是：

$$S_e^2 = \frac{1}{n-r-1} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_r x_{ir} \varepsilon_i))^2$$

而矩阵形式计算的方差是：（证明请见附录5.1.18）

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

2.8 判断系数的存在性

有的时候部分系数很接近零，我们可以简单地抛弃这些系数不要。怎么决定系数的去留呢？我们之前介绍了和 β_i 有关的假设检验，我们把对应的 β_i 值直接和零做假设检验，如果概率不小于显著性水平，我们就认为这个对应的 β_i 等于零，简化我们的模型。

3 评估、诊断模型并对其进行改进

3.1 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）。

3.2 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。

3.3 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。

4 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。

4.1 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。

4.1.1 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。

4.1.2 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。

4.1.3 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。

4.2 单因素方差分析在线性回归中的应用

4.3 双因素方差分析

5 附录 Appendix

5.1 证明

5.1.1 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。

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

所以函数在 $\nu = \bar{u}$ 时取得最小值，证毕。

5.1.2 最小二乘法参数的计算

根据最小二乘法的定义（2.2），我们知道：

$$\begin{aligned} Q(\beta_0,\beta_1) &= \sum_{i=1}^{n}(y_i-(\beta_0+\beta_1 x_i))^2 \\ &= \sum_{i=1}^{n}((y_i-\beta_1 x_i)-\beta_0)^2 \end{aligned}$$

所以根据上面的证明（5.1.1），要使得 Q 最小，有 $\hat{\beta}_0=\overline{(y_i-\beta_1 x_i)}=\bar{y}-\beta_1\bar{x}$，得到 $\hat{\beta}_0$ 的计算公式。接下来将得到的最优 $\hat{\beta}_0$ 代入公式中，得到：

$$\begin{aligned} Q(\hat{\beta}_0,\beta_1) &= \sum_{i=1}^{n}(y_i-(\hat{\beta}_0+\beta_1 x_i))^2 \\ &= \sum_{i=1}^{n}[y_i-(\bar{y}-\beta_1\bar{x}+\beta_1 x_i)]^2 \\ &= \sum_{i=1}^{n}[(y_i-\bar{y})-(x_i-\bar{x})\beta_1]^2 \end{aligned}$$

然后我们证明对一列实数 $\{u_i\}$ 和 $\{a_i\}$，函数 $q(\nu)=\sum_{i=1}^{n}(u_i-a_i\nu)^2$，它取得最小值当且仅当 $\nu=\frac{\sum_{i=1}^{n}a_iu_i}{\sum_{i=1}^{n}a_i^2}$：

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

易知它是开口向上的抛物线，最小值仅在对称轴处取得，此抛物线的对称轴 $-\frac{b}{2a}=-\frac{-2\sum_{i=1}^{n}a_iu_i}{2\sum_{i=1}^{n}a_i^2}=\frac{\sum_{i=1}^{n}a_iu_i}{\sum_{i=1}^{n}a_i^2}$。将这个结论用于得到的 $Q(\hat{\beta}_0,\beta_1)$ 结果，我们能知道只有当 $\beta_1=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$ 时，$Q(\hat{\beta}_0,\beta_1)$ 能取得最小值，所以 $\hat{\beta}_1=\frac{S_{xy}}{S_{xx}}$，证毕。

5.1.3 残差方差的估计

（此处证明并不完全）

我们一直使用最小二乘法得出的 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 用作最佳参数。现在我们认为这两个参数存在真值 β_0 和 β_1，假若它们的值已知，则有 $\varepsilon_i=Y_i-(\beta_0+\beta_1x_i)\sim N(0,\sigma^2)$，于是我们的样本方差就是总体方差的**无偏估计**，即 $\frac{1}{n}\sum_{i=1}^{n}[(Y_i-(\beta_0+\beta_1x_i))^2]$ 就可以用于估计 σ^2。

但是在实际操作中，我们总是用 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 来估计真值，这就引入了两个“参数”。自由度的大小总是样本量减去参数个数，所以此处我们将分母替换成 $n-2$，以解决此问题。

5.1.4 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 $\hat{\beta}_1=\dfrac{S_{xy}}{S_{xx}}$，所以 $\hat{\beta}_1 S_{xy}=\dfrac{S_{xy}^2}{S_{xx}}=\hat{\beta}_1^2 S_{xx}$，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。

5.1.5 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。

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

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。

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

5.1.6 最小二乘法系数的协方差证明

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

5.1.7 最小二乘法系数分布的证明 1

（此证明暂时不要求掌握）

5.1.8 最小二乘法系数分布的证明 2

（此证明暂时不要求掌握）

5.1.9 预测值分布的证明 1

5.1.10 预测值分布的证明 2

5.1.11 预测值分布的证明 3

5.1.12 预测区间和置信区间的不同

5.1.13 残差满足的性质

5.1.14 残差的标准化与学生化

（此证明不要求掌握）

5.1.15 列向量必须线性无关

5.1.16 列线性无关的矩阵性质 1

5.1.17 多元最小二乘法系数的计算

5.1.18 多元最小二乘法残差的计算