时间序列分析大作业

课题题目： 最小二乘法的贝叶斯方法解释

所在学院： 数学与统计学院

小组成员： 李海豹 唐超 侯同帅 翟泓智 徐少波

任课教师： 王晓

年　　月　　日

最小二乘法的贝叶斯方法解释

一、背景与意义

最小二乘法是提供“观测组合”的主要工具之一,它依据对某事件的大量观测而获得“最佳”结果或“最可能”表现形式。如已知两变量为线性关系$y = ax + b$,对其进行$n\ (n > 2)$次观测而获得n 对数据。若将这n对数据代入方程求解a,b之值则无确定解。最小二乘法提供了一个求解方法，其基本思想就是寻找“最接近”这n个观测点的直线。

美国统计学家斯蒂格勒(S. M. Stigler)说,“最小二乘法之于数理统计学犹如微积分之于数学”。由此可见最小二乘方法在统计学上的地位之高!

然而我们对最小二乘的理解仅仅停留在应用上，却不知道为什么用最小二乘法，也不知道最小二乘法的前提高斯-马尔科夫假定为什么合理。因此，我们打算从误差分布、最小二乘的贝叶斯方法解释两个方面去探究最小二乘。

二、贝叶斯和高斯简介

1、 贝叶斯

贝叶斯(Thomas Bayes,1702—1761)英国牧师、业余数学家。为了证明上帝的存在，他发明了概率统计学原理，遗憾的是，他的这一美好愿望至死也未能实现。贝叶斯在数学方面主要研究概率论。他首先将归纳推理法用于概率论基础理论，并创立了贝叶斯统计理论，对于统计决策函数、统计推断、统计的估算等做出了贡献。1763 年发表了这方面的论著，对于现代概率论和数理统计都有很重要的作用。贝叶斯的另一著作《机会的学说概论》发表于 1758 年。

贝叶斯所采用的许多术语被沿用至今。贝叶斯思想和方法对概率统计的发展产生了深远的影响。今天，贝叶斯思想和方法在许多领域都获得了广泛的应用。从二十世纪 20~30 年代开始,**概率统计学出现了“频率学派”和“贝叶斯学派”的争论**，至今，两派的恩恩怨怨仍在继续。

2、 高斯

约翰·卡尔·弗里德里希·高斯（德语：Johann Karl Friedrich Gauß ，1777 年 4 月 30 日－1855 年 2 月 23 日）， 德国数学家、物理学家、天文学家、

大地测量家，生于布伦瑞克，卒于哥廷根。高斯被认为是历史上最重要的数学家之一，并有“**数学王子**”的美誉。

高斯 9 岁时，用很短的时间计算出了小学老师布置的任务：对自然数从 1 到 100 的求和。他所使用的方法是：对 50 对构造成和 101 的数列求和（1+100, 2+99…, 50+51），同时得到结果：5050。

高斯 12 岁时，已经开始怀疑几何原本中的基础证明。

当他 16 岁时，预测在欧氏几何之外必然会产生一门完全不同的几何学，即非欧几里德几何学。后来，相对论证明了宇宙空间实际上是非欧几何的空间。高斯的思想被近 100 年后的物理学接受了。

高斯 19 岁时，在大学仅用尺规便构造出了 17 边形。

高斯总结了复数的应用，并且严格证明了每一个n阶的代数方程必有n个实数或者复数解。

1840 年，他和韦伯画出了世界第一张地球磁场图，并且定出了地球磁南极和磁北极的位置。

高斯在最小二乘法基础上创立的测量平差理论的帮助下，测算出了小行星谷神星的运行轨迹。

三、 贝叶斯公式与贝叶斯判别

条件概率：

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

于是得到：

$$P(AB) = P(A|B)P(B) = P(B|A)P(A)$$

当条件B不单一的时候，我们假设Ω是一个完备事件组，即：

$$\Omega = \{B_1, B_2, \cdots, B_n\}$$

且： $B_i \cap B_j = \emptyset$ ， $i \neq j$， $i, j = 1,2, \ldots, n$

$$\mathrm{A} = \mathrm{A}\Omega = \mathrm{A}B_1 + AB_2 + \cdots + AB_n$$

$AB_1, AB_2, \ldots, AB_n$也相互独立，自然下面的式子也成立：

$$\mathrm{P}(\mathrm{A}) = \mathrm{P}(\mathrm{A}\Omega) = \mathrm{P}(\mathrm{A}B_1) + P(AB_2) + \cdots + P(AB_n)$$

应用条件概率公式有：

$$P(A) = \mathrm{P}(A|B_1)P(B_1) + \cdots + P(A|B_n)P(B_n) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

这就得到了**全概率公式**。

有了全概率和条件概率的基础，进一步我们有：

$$P(AB_j) = P(B_j|A)P(A) = P(B_j|A)\sum_{i=1}^{n} P(A|B_i)P(B_i)$$

即：

$$P(B_j|A) = \frac{P(\mathrm{A}|\mathrm{B}_j)P(\mathrm{B}_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$$

这就是**贝叶斯公式的**最终样貌，到此，贝叶斯公式推导完成。

由于后面工作的需要，我们在这里对**贝叶斯判别准则**进行导出。

$G_1, G_2, \cdots, G_k$，为k个p维总体，分别具有概率密度函数：$f_1(x), f_2(x), \cdots, f_k(x)$。各总体的先验概率分别为：

$$p_1 = P(G_1), p_2 = P(G_2), \cdots, p_k = P(G_k)$$

并且有：

$$p_1 + p_2 + \cdots + p_k = 1$$

一般地，当我们对总体信息了解不足时，会作出以下假设：

$$p_1 = p_2 = \cdots = p_k = \frac{1}{k}。$$

对于一个新的样本$x = (x_1, x_2, \cdots, x_p)^T$，那么，它属于某一总体$G_k$的后验概率为：

$$P(G_i|x) = \frac{P(G_i)f_i(x)}{\sum P(G_i)f_i(x)} = \frac{p_i f_i(x)}{\sum p_i f_i(x)}$$

于是，我们得到**贝叶斯判别准则**：若

$$P(G_i|x) = \underbrace{max}_{1\le j\le k}\{P(G_j|x)\}，\ i = 1,2,\cdots,k$$

则判样本$x \in G_i$。

注：当达到最大后验概率的G_i不止一个时，该方法失效，样本待判。

四、 高斯误差分布函数

为什么是高斯误差分布函数？

其实，早在高斯之前，拉普拉斯也做了和高斯一样的事情，最终得了到**拉普拉斯分布函数：**

$$f(x)=\frac{m}{2}e^{-m|x|},x\in(-\infty,+\infty)$$

然而，拉普拉斯很快发现，基于这个误差函数进行的计算是相当繁杂的，故也就不可能有多大的实际应用价值。后来，拉普拉斯得到一个更加复杂的函数表达式，只好无功而返了。

另外一个人——棣莫弗，通过二项分布的近似公式也导出了正态分布函数，但它并没有将其应用到误差分布上来，因此与这一重大发现擦肩而过。

在推导高斯误差密度函数之前，我们需要接受以下假设与引理：

拉普拉斯误差分布条件：

① $f(x)=f(-x)$；

② $x\to\infty$时,$f(x)\to 0$(因无限大误差的概率为0)；

③ $\int_{-\infty}^{\infty}f(x)dx=1$(因在任意两数值之间曲线下方的面积代表观测具有的误差在这两个值之间的概率)。

引理1：

若函数$g(x)$为具有二阶导数的偶函数,则$g'(x)$是奇函数,$g''(x)$又是偶函数。

引理2：

若函数$g(x)$满足以下条件：

① $g(0)=0$；

② $g(x)$可导且导函数连续；

③ $g'(x)$是偶函数；

④ 对任意自然数m及实数x满足：$g'(mx)=g'(x)$;

则对任意实数x，函数$g(x)$必具有形式：

$$g(x)=cx\text{(其中}c\text{为常数)}$$

引理3：

函数e^{-x^2}在整个实数域上的积分值为$\sqrt{\pi}$，即：

$$\int_{-\infty}^{\infty}e^{-x^2}\,dx=\sqrt{\pi}\quad\text{（可利用积分套证明）}$$

做完准备工作，下面开始误差分布函数的推导。

设误差密度函数为$f(x)$，总体真值为θ，一个关于测量误差的简单随机样本$X_1,X_2,\cdots,X_n$，考虑使如下似然函数取最大值时的估计量：

$$\mathrm{L}(\hat{\theta})=\underset{\theta}{max}L(\theta)=\underset{\theta}{max}\{f(x_1-\theta)f(x_2-\theta)\cdots f(x_n-\theta)\}$$

与传统的极大似然估计不同——**在已知分布函数的情形下，求使得上式达到最大的$\hat{\theta}$**，高斯的创新性思维在于，既然经验已经告诉我们样本的算数平均值是使得上式达到最大的优良估计，那么是什么样的分布函数导致的？即**高斯先承认了θ的最优估计为$\bar{x}$。**

对似然函数取对数，则原先的问题等价于：

$$maxL(\hat{\theta})\Leftrightarrow max\sum_{i=1}^{n}lnf(x_i-\hat{\theta})$$

要使得上式取最大值，必然有：

$$\sum_{i=1}^{n}\left(lnf(x_i-\hat{\theta})\right)'=\sum_{i=1}^{n}\frac{f'(x_i-\hat{\theta})}{f(x_i-\hat{\theta})}=0$$

构造辅助函数：

$$g(x_i-\hat{\theta})=\frac{f'(x_i-\hat{\theta})}{f(x_i-\hat{\theta})}$$

则：

$$\sum_{i=1}^{n}g(x_i-\hat{\theta})=0$$

为得到$g(x)$的形式，利用f(x)的偶函数性质并运用引理1的结论知，$g(x)$是奇函数。所以有：

$$g(x)=-g(x),g(0)=0$$

取自然数m，并令$n=m+1$，$x_1=x_2=\cdots=x_m=-x$，$x_{m+1}=mx$，

则此时$\hat{\theta}=\bar{x}=0$

由于：$\sum g(x_i-\hat{\theta})=\sum g(x_i)=0$

即：

$$\sum g(x_i)=\underbrace{g(-x)+\cdots+g(-x)}_{m}+g(mx)=mg(-x)+g(mx)=0$$

又由于g(x)是奇函数，当然有：$\boldsymbol{mg(-x)=-mg(x)}$

从而得到：$\boldsymbol{g(mx)=mg(x)}$

上式对一切自然数m及实数x成立。由此，假定$g(x)$可导且导函数连续，等式两边分别关于x求导，得：

$$\frac{dg(mx)}{d(mx)}\frac{d(mx)}{dx}=m\frac{dg(x)}{dx}$$

即：$\boldsymbol{g'(mx)=g'(x)}$

由于$g'(x)=\frac{f''(x)f(x)-[f'(x)]^2}{f^2(x)}$，而$f(x)$是偶函数，根据引理1，$f'(x)$是奇函数，$f''(X)$是偶函数，于是$g'(x)$是偶函数。由于$f(x)$具有二阶连续导函数，所以$g'(x)$连续。这样，$g(x)$符合引理2的全部条件，运用其结论得：

$$g(x)=\frac{f'(x)}{f(x)}=cx$$

两边积分可得：

$$\int\frac{f'(x)}{f(x)}dx=\int\frac{df(x)}{f(x)}=\frac{1}{2}cx^2+c'$$

即：$ln\,f(x)=\frac{1}{2}cx^2+c'$

从而:

$$f(x)=Me^{\frac{1}{2}cx^2},\ (M=e^{c'})$$

上述$f(x)$显然恒大于零，要使其能成为密度函数还须使其在整个实数域上积分值为1。于是有：

$$\int_{-\infty}^{+\infty}f(x)dx=\int_{-\infty}^{+\infty}Me^{\frac{1}{2}cx^2}dx=1$$

显然，c必须为小于零的常数，记$c=-\frac{1}{\sigma^2}$，上式变为：

$$\int_{-\infty}^{+\infty}Me^{-\frac{1}{2\sigma^2}x^2}dx=1$$

运用引理3的结论：

$$\int_{-\infty}^{+\infty}e^{-x^2}dx=\sqrt{\pi}$$

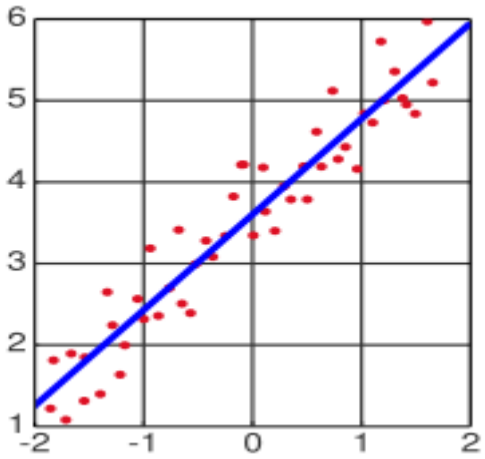
令$y=\frac{1}{\sqrt{2}\sigma}x$，可以得到$M=\frac{1}{\sqrt{2\pi}\sigma}$

从而高斯误差分布函数具有形式：

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}x^2}$$

到此，高斯误差分布函数也已经推导完成（详情参见《对高斯分布函数形式的推导》）。

五、 贝叶斯与最小二乘



我们都知道使用经典的最小二乘方法来做线性回归。问题描述是：给定平面上N个点（假设是线性关系的数据），找出一条最佳描述了这些点的直线。

问题是，我们如何定义最佳？

设每个点的坐标为(X_i,Y_i)，如果回归直线为$y=f(x)$。那么(X_i,Y_i)与该直线对这个点的“预测”$(X_i,f(X_i))$就相差了：$\Delta Y_i=|Y_i-f(X_i)|$。所有的这些误差的平方的和我们记为Rss，则：

$$Rss=\sum {\Delta Y_i}^2$$

最小二乘就是要寻找使得$Rss=\sum {\Delta Y_i}^2$最小，即当$Rss=min\sum {\Delta Y_i}^2$时对应的那条直线。

至于为什么是误差的平方和而不是误差的绝对值和，统计学上也没有什么好的解释。然而贝叶斯方法却能对此给出一个完美的解释。

我们假设直线对于X_i给出的最有可能的预测是$f(X_i)$，所有偏离$(X_i,f(X_i))$的数据点都是受到了噪音干扰，才使得它们偏离了完美的一条直线。前面我们已

经推导了误差是服从高斯分布的，这个分布曲线以$(X_i, f(X_i))$中心，实际纵坐标为Y_i的点(X_i, Y_i)发生的概率为：

$p = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(Y_i - f(X_i))^2}{2\sigma^2}}$，正比于$e^{-\Delta {Y_i}^2}$。

记

A：已观测到的N个数据点；

用A_i，$i = 1,2,\cdots,n$，表示A中的每个数据点。

B_j：我们要寻找的直线为$f_j(x)$；

由于待选的可能直线有很多，且优良性未知，一般我们设每条直线被选择的先验概率相同（贝叶斯判别假设），并且B_j，$j = 1,2,\cdots,n$，构成一个完备事件组。（为了方便，下面直接使用B而不使用B_j）

现在问题又回到贝叶斯方面，我们要想最大化的后验概率：

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

A为观测量，即$P(A)$是一个定值，那么：

$$P(B|A) \propto P(A|B)P(B)$$

我们已经假定每一个$P(B)$均相等, 为一定值，故我们只需要关注$P(A|B)$，这一项是这些数据点在直线$f(x)$上的概率。

而各数据点之间两两独立，故他们都在直线$f(x)$上的概率为：

$$P(B|A) = \prod_{i=1}^{n} P(B|A_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(Y_i - f(X_i))^2}{2\sigma^2}} \propto \prod_{i=1}^{n} e^{-\Delta {Y_i}^2}$$

即：$P(B|A) \propto \prod_{i=1}^{n} e^{-\Delta Y_i{}^2} = e^{-\sum \Delta Y_i{}^2}$

进一步有：

$$max\, P(B|A) = min \sum \Delta Y_i{}^2$$

于是，最大化$P(B|A)$，就转化成了最小化$\sum \Delta Y_i{}^2$的问题,显然，这就是最小二乘的形式，这样，最小二乘法在贝叶斯方法上就得到了完美的解释。

参考文献：

[1] 尉迟江. 对高斯分布函数形式的推导[J]. 统计与信息论坛, 2009(5):3-6.

[2] 贾小勇, 徐传胜, 白欣. 最小二乘法的创立及其思想方法[J]. 西北大学学报(自然科学版), 2006(3).

[3] 刘乐平等. 贝叶斯身世之谜——写在贝叶斯定理发表 250 周年之际[J]. 统计研究, 2013, 30(12):3-9.

[4] R 语言中文社区. 上帝手中的骰子——无所不能的贝叶斯(上篇). CSDN, 2018.

[5] zwan0518. 贝叶斯从浅入深详细解析，详细例子解释. CSDN, 2013.