各位旅客朋友，您好，欢迎您乘坐椒盐海豹号列车，本次列车由线性回归开往 IRT 模型，请您上车后按照车票上指定的车厢和座位号对号入座。本次列车始发站为线性回归，途径变换后可化为线性回归的非线性回归、Logistic 回归、广义线性模型、隐变量与 EM 方法，最终到达终点站 IRT 模型。

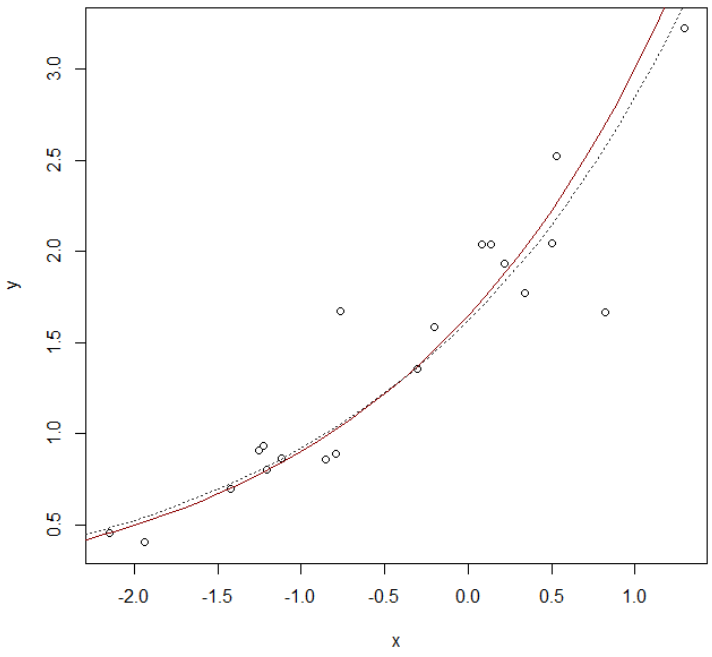
我们假设您已经具备了截至线性回归（包括线性回归）的基础统计学知识，包括但不限于对线性回归有基本了解，大致了解常见的数学符号的记法（比如∑表示累加，Π 表示累乘），了解概率、似然、概率密度函数等基础概念，了解各个维度彼此独立的多维随机变量的联合概率密度与边际概率密度之间的关系（其实就是乘一起的关系）等。【当然如果对这些概念不了解也不要紧，看到不懂的地方度娘一下就可以】

本次列车马上就要开车了，请您再次检查自己的车票是否与本次列车相符。

1. 先从半对数回归说起

在回归分析中，我们并不总是能够有足够的好运面对线性关系的数据，例如下列数据对：

x	y
-0.850	0.862
-1.207	0.806
0.216	1.933
1.294	3.223
-0.203	1.585
-0.763	1.675
-1.115	0.868
0.342	1.772
-1.254	0.909
0.138	2.042
-1.937	0.403
-1.223	0.932
0.081	2.039
0.532	2.527
0.501	2.047
-2.150	0.456
-0.307	1.357
-0.793	0.890
-1.423	0.696
0.820	1.664



通过绘制散点图，不难发现 x 与 y 之间可能存在指数关系（即$y = e^{ax+b}$）。对于这类数据，通常的做法是通过一些手段对变量进行“变换”，使得变换后的结果能够适用于线性回归。

对于指数型数据而言，通常采用“对数变换”的方式将其转化为线性回归（注意到，对$y = e^{ax+b}$两边取对数，容易有$\ln y = ax + b$，如果令$z = \ln y$，则 z 与 x 之间呈线性关系）。对于我们给出的数据对，对 y 进行对数变换后再进行回归，可以求得$y = e^{0.5629x+0.4827}$，结果和生成数据所用的模型（$y = e^{0.6x+0.5}$）还是十分接近的（生成数据的模型为图中的红色实线，回归模型为图中的灰色虚线）。

事实上，对许多“通过变形可以将其化为线性回归”的问题而言，选择合适的变换手段将其转化为线性回归问题是解决这类问题的“万金油”。而在日常实践中，对于一些分布具有强理论假定的变量（特别是一些人口学变量）数据变换也是研究者常用的纠偏手段（方法不限于对数变换，平方根变换和角变换也是常用的手段，虽然事实上这种转换往往是将不符合正态分布的数据变成了另外一些奇形怪状的分布，但在大多数情况下这种变换总能将不可接受的误差减小到一个可接受的范围内）。

2. 线性模型的极大似然估计

之前的一节大致讲述了 x 与 y 的关系非线性关系的情况下研究者可以通过选择合适的函数对 y 进行恰当的变换，从而使得变换后的 y，即 f(y)与 x 之间呈线性关系，进而使用传统的线性回归方法对该数据进行分析。这种转化的思路解决了一些问题，但对一些更特殊的情况而言，找到直接对原始数据 y 进行变换的方法仍然是困难的，例如下述问题：

例：某高校心理学兴趣小组希望探究个体的知识水平 x 与他们能否正确回答某题目（解答题）之间的关系，已知该小组测量出的个体知识水平 x 是连续变量，我们约定如果个体 i 正确回答了该题目，则记 $y_i = 1$，否则 $y_i = 0$，试讨论 x 与 y 之间的关系。

x	-0.312	0.368	-0.593	0.830	0.471	0.859	1.067	-0.011	-0.412	0.700	……
y	0	1	0	1	1	1	1	0	0	1	……

示例：一种可能情况下的 x，y 组合

注意到这里的 y 只有 0 和 1 两种取值，这意味着我们之前介绍的对 y 变换的方法变得不再适用，这种情况下应该怎么办呢？

别急，让我们回到最正统的线性回归问题，假定数据符合 Gauss-Markov 假设，并且误差项服从正态分布。容易知道，对于线性回归模型$y = ax + b$而言，当我们已知$x = x_0$时，y 服从均值为$ax_0 + b$，标准差为 σ 的正态分布，此时容易有$y = y_0$的可能性（也可以近似把它理解成“概率”）为：

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{[y_0-(ax_0+b)]^2}{2\sigma^2}}$$

那么，如果我们目前手头有一个简单随机样本(x_1, y_1), (x_2, y_2), …, (x_n, y_n)，并且我们假定 x 与 y 之间存在有$y = ax + b$的关系，那么很容易可以计算出当 x 为 $x_1, x_2, \ldots, x_n$ 时，对应的 y 分别为 $y_1, y_2, \ldots, y_n$ 的可能性为：

$$L(a,b) = P(y|x;a,b) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{[y_i-(ax_i+b)]^2}{2\sigma^2}}$$

容易发现，在给定样本的情况下，上式是一个关于a和b的函数。进一步，求出当上式取最大值时的(a,b)，这样的(a,b)就可以作为参数的**极大似然估计**。用通俗的话来讲，总体参数的极大似然估计求的就是“使得出现既定样本的可能性最大的参数”。直接求上面这个式子的最大值好像还有些难求，但好在可以通过适当的变形，使变形后的函数和原函数的极大值点位置相同，对于似然函数来说，通常采用的变形方式是直接取对数（这样可以在不改变驻点位置的情况下把累乘变为累加，相对累乘而言，累加就很好处理了）。变形后的函数如下：

$$\ln L(a,b) = \ln \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{[y_i-(ax_i+b)]^2}{2\sigma^2}} = \sum_{i=1}^{n} \ln \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{[y_i-(ax_i+b)]^2}{2\sigma^2}}$$

$$= C - k\sum_{i=1}^{n} [y_i - (ax_i + b)]^2$$

由于方差的非负性，容易知道 k 显然是一个正数，此时容易发现，最大化$\ln L(a,b)$等价于最小化$[y_i - (ax_i + b)]^2$，容易发现，这和最小二乘法的优化目标是一致的。事实上，极大似然估计给我们提供了一个解决问题的新思路——当我们不能通过对原始数据进行简单变换的方式解决问题时，如果我们有一个明确的理论模型指向已知自变量时因变量的分布，那么我们就可以借助极大似然法求取出模型中的参数。

听上去还是很抽象？不要紧，我们来借助这个视角重新审视半对数回归。半对数回归的因变量在已知自变量的情况下服从位置参数为$ax_0 + b$，尺度参数为σ的对数正态分布，即对模型$y = e^{ax+b}$而言，当已知$x = x_0$时，有 y 的概率密度为：

$$\frac{1}{y\sqrt{2\pi}\sigma} e^{-\frac{[\ln y-(ax_0+b)]^2}{2\sigma^2}}$$

进而可以写出此时的对数似然函数为：

$$\ln L(a,b) = \ln \prod_{i=1}^{n} \frac{1}{y_i\sqrt{2\pi}\sigma} e^{-\frac{[\ln y_i-(ax_i+b)]^2}{2\sigma^2}} = C - k\sum_{i=1}^{n}[\ln y_i - (ax_i+b)]^2$$

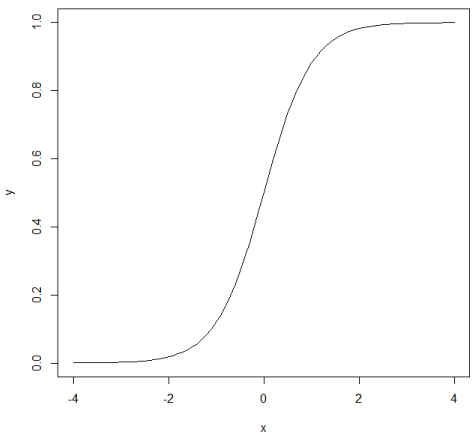
容易发现，此时最大化$\ln L(a,b)$等价于最小化$[\ln y_i - (ax_i+b)]^2$，这相当于对 y 做变形后对数据使用最小二乘法得到的结果。

3. logistic 回归

再次回到我们的例子中，有了前面的思路，我们这里就可以“照猫画虎”解决例子中的问题，我们不妨假设 y 遵循这样的分布：当已知$x = x_i$时，y_i以概率$ax_i + b$等于 1，以概率$1 - (ax_i + b)$等于 0。这样似乎就能把问题划归为已知问题了。

等等，事情真的这么简单么？

注意到，我们这里以一个线性函数（$ax + b$）模拟事件发生的概率，但小学二年级的学生都知道，概率应当是一个处于[0, 1]之间的值，而线性函数是一个无界函数，显然不可能满足这个需求。这个时候要怎么办呢？这时候有聪明的小伙伴就会想到，用一个有界函数建立R→[0, 1]的映射不就可以了嘛！实际上也的确是这样做的，采用的“桥”就是下面这个东西：



其函数表达式为：

$$\frac{1}{1+e^{-x}}$$

这个函数是标准 logistic 分布的累积分布函数。在能将全体实数映射到[0, 1]区间的同时，选用该函数还赋予了模型更实际的物理意义。

※ 考虑个体的听感觉加工问题，假定个体接收到一个强度为 t 的声音信号，并且该信号在个体的听觉系统中加工后的信号 t’，其强度服从 logistic 分布，即$t' \sim L(\mu_i, \gamma_i)$，假定个体的决策标准为 c_i，即只有当处理后的信号 t’的强度大于 t_c 时，个体才会确定自己听到了声音，那么容易给出个体确定自己听到声音的概率为：

$$\int_{t=c_i}^{+\infty} \frac{e^{-\frac{x-\mu_i}{\gamma_i}}}{\gamma\left(1+e^{-\frac{x-\mu_i}{\gamma_i}}\right)^2} dx = 1 - \frac{1}{1+e^{-\frac{c_i-\mu_i}{\gamma_i}}} = \frac{1}{1+e^{\frac{c_i-\mu_i}{\gamma_i}}}$$

不难发现，当给予一个声音刺激时，个体确定自己听到它的概率只与个体的决策标准以及个体的感觉系统的处理功能有关，如果我们将这些内容“打包”成一个参数 θ_i，即$\theta_i = \frac{\mu_i - c_i}{\gamma_i}$时，并将其命名为“个体能力”时，容易有对于某刺激，个体给出“是”这个反应的可能为：

$$\frac{1}{1+e^{-\theta_i}}$$

可能有好奇的同鞋会问“为什么要假定是 logistic 分布而不假定是正态分布啊？”事实上假定为正态分布也是完全可以的，只不过 logistic 分布的累积分布函数更好算一点，并且这两个分布的差异不是很大，所以实践中更多使用 logistic 分布的累积分布（事实上利用正态分布的累积分布函数来建立全体实数到[0, 1]区间映射的方法叫做 Probit 回归/概率单位回归）。

借助 logistic 分布累积分布函数的这座“桥”，我们容易建立下述 y 的分布：当已知$x = x_i$时，y_i以概率$\frac{1}{1+\exp[-(ax_i+b)]}$等于 1，以概率$\frac{1}{1+\exp(ax_i+b)}$等于 0，此时，根据前面讲过的例子，可以写出 logistic 模型的对数似然函数为：

$$\ln L(a,b) = \ln \prod_{i=1}^{n} \frac{1}{1+\exp[-(ax_i+b)]}^{y_i} \frac{1}{1+\exp(ax_i+b)}^{1-y_i}$$

$$= \ln \prod_{i=1}^{n} \frac{1}{\{1+\exp[-(ax_i+b)]\}^{y_i}} \frac{\exp[-(ax_i+b)]^{1-y_i}}{\{1+\exp[-(ax_i+b)]\}^{1-y_i}}$$

$$= \ln \prod_{i=1}^{n} \frac{\exp(ax_i+b)^{y_i}}{1+\exp(ax_i+b)} = \sum_{i=1}^{n} \{y_i(ax_i+b) - \ln[1+\exp(ax_i+b)]\}$$

*特别地，注意到当我们假设 y_i 以概率$f(x)$等于 1，以$1-f(x)$等于 0 时，如果我们将 yi 视为我们需要贴近的真实数据，将概率$f(x)$视作模型给出的模拟值，根据交叉熵损失函数的定义，容易有：

$$Loss_{CE} = -\frac{1}{n}\sum_{i=1}^{n} \{y_i \ln f(x_i) + (1-y_i)\ln[1-f(x_i)]\}$$

容易发现在 logistic 回归中使似然函数极大等价于使交叉熵损失函数最小。

与普通的最小二乘法不同的是，logistic 回归的对数似然函数的极值难以求出一个明确的解析解（说白了就是不像最小二乘法那样可以给出一个明确的公式快速计算参数）。在实际求取对数似然函数极大值时，一般会使用牛顿法或者梯度上升法求取似然函数极大值的数值解。

关于 logistic 回归的一个有趣的结论是，logistic 回归的系数与 OR 值具有天然的联系。注意到 OR 值的定义为：

$$OR = \frac{\frac{P(Y=1|X_1)}{P(Y=0|X_1)}}{\frac{P(Y=1|X_0)}{P(Y=0|X_0)}} = \frac{\frac{1+\exp(ax_1+b)}{1+\exp[-(ax_1+b)]}}{\frac{1+\exp(ax_0+b)}{1+\exp[-(ax_0+b)]}} = \frac{\exp(ax_1+b)}{\exp(ax_0+b)} = \exp[a(x_1-x_0)]$$

当我们令 x_1 与 x_0 相差一个单位（对于连续变量）或 x_1 代表被试具有某属性 x_0 代表被试不具有某属性（对于二分变量）时，容易有$\exp[a(x_1-x_0)] = \exp a$。换言之，logistic 回归的偏回归系数是控制了其他因素情况下单独使某个变量变化一个单位时比值比的对数。当回归系数大于 0 时，对应的 OR 值大于 1，说明对应的因素更可能引起 y 取 1 时所代表的结果，若回归系数小于 0 时，对应的 OR 值小于 1，说明对应的因素更可能引起 y 取 0 时所代表的结果。若回归系数为 0，对于的 OR 值等于 1，说明对应因素与 y 没关系。

4. 从 logistic 回归说开去：广义线性模型、隐变量与期望最大化方法

现在，让我们从另一个角度重新理解前面提及的 logistic 回归以及半对数回归。我们假定我们收集到的数据(x_i, y_i)都是不完整的数据，中间遗漏了一个变量 z_i。换言之，实际的数

据集是(x_i, y_i, z_i)，但由于种种原因，我们只收集到了(x_i, y_i)，但我们大概知晓下面两个信息：

① 当$x = x_i$时变量z的分布$f(z|x_i)$

② 当$z = z_i$时变量y的分布$f(y|z_i)$

此时，可以借助变量 z 重写原对数似然函数：

$$\ln L(a,b) = \ln \prod_{i=1}^{n} p(y_i|x_i;a,b) = \ln \prod_{i=1}^{n} \int_{-\infty}^{+\infty} p(y_i, z_i|x_i;a,b)\mathrm{d}z_i$$

$$= \sum_{i=1}^{n} \ln \int_{-\infty}^{+\infty} p(y_i|z_i, x_i;a,b)p(z_i|x_i;a,b)\mathrm{d}z_i$$

容易知道，半对数回归的情况相当于：

$$p(z_i|x_i;a,b) = \frac{1}{\sqrt{2\pi}\sigma}\exp -\frac{(z_i - ax_i - b)^2}{2\sigma^2}$$

$$p(y_i|z_i, x_i;a,b) = \begin{cases} 1, & y_i = \exp z_i \\ 0, & otherwise \end{cases}$$

而对于 logistic 回归的情况，则有：

$$p(z_i|x_i;a,b) = \begin{cases} 1, & z_i = ax_i + b \\ 0, & otherwise \end{cases}$$

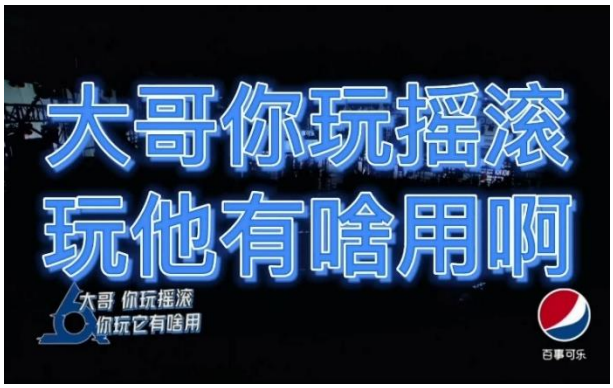
$$p(y_i|z_i, x_i;a,b) = \frac{\exp y_i z_i}{1 + \exp z_i}$$

不难发现，对于半对数回归和 logistic 回归而言，模型中都包括一个确定性部分和一个随机部分。只不过对半对数回归（以及此类通过变换 y 来将问题转化为线性回归的模型）而言，确定性部分在隐变量 z 与因变量 y 之间，所以可以通过对因变量 y 进行逆变换直接求出隐变量 z，进而建模求解。而对于 logistic 回归而言，确定性部分在自变量 x 与隐变量 z 之间，因此需要通过一些“特殊手段”建立起 y 与 z 的关系，通常是假定 y 的分布的参数与隐变量 z 的函数之间存在某种关系。需要注意的是，由于隐变量 z 是通过自变量 x 的线性变换得到，而 y 的分布参数往往具有某种性质，故实际应用时需要根据变换前后的性质选择合适的“连接函数”，例如：

y 的分布	参数及其特征	与隐变量的连接方式	对应的模型
两点分布	$\theta \in [0,1]$	$\theta = \frac{1}{1+\exp -z}$	Logistic 回归
两点分布	$\theta \in [0,1]$	$\theta = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}\exp -\frac{t^2}{2}\mathrm{d}t$	Probit 回归
泊松分布	$\lambda > 0$	$\lambda = \exp z$	Possion 回归

这种将线性模型生成的隐变量通过特定函数与某分布（通常是指数族分布）中的参数建立起

联系，进而对数据建模的方法，就是传说中的“广义线性模型”，上表中举出的三个例子则是广义线性模型中常见的几类具体的模型。看到这里，有的同学可能又会疑惑了，虽然我们之前大费周章地介绍了隐变量，但实际上不管是广义线性模型，还是变量转换，实际上都有一部分是“确定”的，在这种情形下，很容易写出似然函数进而求出对应的参数，完全用不到“舍近求远”来引入隐变量，那引入隐变量到底有啥用呢？



让我们在前面关于 logistic 回归的例题上更进一步：

例子升级版：该心理学兴趣小组在进一步收集资料的过程中发现，既有的用来测量知识水平（x）的工具都不够理想。于是，该小组希望引入新题目，通过考察被试在多道同类题目上的作答来估计不同题目在面对不同知识水平被试时的表现。设该小组最后开发出了 m 道题，我们仍然约定如果个体 i 正确回答了题目 j，则记 $y_{ij}=1$，否则 $y_{ij}=0$。假定个体 i 的知识水平 x_i 服从正态分布，已知当个体知识水平为 x_i 时，个体正确回答第 j 题的概率为：

$$P(y_{ij}=1|x_i;a_j,b_j)=\frac{1}{1+\exp-(a_jx_i+b_j)}$$

其中 a_j，b_j 是只与题目有关，与被试无关的参数。试求 a_j 与 b_j 的估计值。

注意到，上例中的似然函数为：

$$\ln L(a,b)=\ln\prod_{i=1}^{n}\prod_{j=1}^{m}p(y_{ij}|a_j,b_j)$$

不难发现，在不借助隐变量（个体知识水平 x）的情况下，直接写出对数似然函数是困难的。但在假设个体 i 的知识水平为 x_i 时，可以借助已知条件写出对数似然函数：

$$\ln L(a,b)=\ln\prod_{i=1}^{n}\prod_{j=1}^{m}\int_{-\infty}^{+\infty}p(y_{ij},x_i|a_j,b_j)\mathrm{d}x_i=\sum_{i=1}^{n}\sum_{j=1}^{m}\ln\int_{-\infty}^{+\infty}p(y_{ij},x_i|a_j,b_j)\mathrm{d}x_i$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{m}\ln\int_{-\infty}^{+\infty}p(y_{ij}|x_i;a_j,b_j)p(x_i|a,b)\mathrm{d}x_i$$

事实上，上面这个似然函数也不是那么容易求出表达式的。（不信的同学可以试试，括弧笑）

这个时候又该怎么办呢？

有小伙伴或许已经看出来了，积分中的$p(x_i|a,b)$式实际上是 x_i 的条件密度，换言之，有$\int_{-\infty}^{+\infty} p(x_i|a,b)\mathrm{d}x_i = 1$，另外，显然地，对数函数是一个凹函数（上凸函数）。提示到这里，高中数学基础比较好的小伙伴可能就能反应过来了，答案是琴生不等式！

[Kono Jensen da!]

琴生不等式给出了上凸函数或下凸函数函数值的期望与期望的函数值之间的关系，即若非负函数$f(x)$满足$\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1$，$g(x)$是可积函数，h 是定义在$g(x)$值域上的上凸函数，那么：

$$\mathrm{h}\left(\int_{-\infty}^{+\infty} g(x)f(x)\mathrm{d}x\right) = \mathrm{h}(E[g(x)]) \geq E(\mathrm{h}[g(x)]) = \int_{-\infty}^{+\infty} \mathrm{h}[g(x)]f(x)\mathrm{d}x$$

回带入对数似然函数，我们有：

$$\ln L(a,b) \geq \sum_{i=1}^{n}\sum_{j=1}^{m}\int_{-\infty}^{+\infty}\left[\ln p\left(y_{ij}|x_i;a_j,b_j\right)\right]p\left(x_i|a_j,b_j\right)\mathrm{d}x_i = E\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\ln p\left(y_{ij}|x_i;a_j,b_j\right)\right]$$

注意到$\sum_{i=1}^{n}\sum_{j=1}^{m}\ln p\left(y_{ij}|x_i;a_j,b_j\right)$实际上是在数据完全收集情况下的对数似然，上式的含义是：考虑隐变量情况下的完全数据对数似然的期望值是实际数据对数似然的下界。故我们可

以进一步通过最大化$\sum_{i=1}^{n}\sum_{j=1}^{m}\int_{-\infty}^{+\infty}[\ln p(y_{ij}|x_i;a_j,b_j)]p(x_i|a_j,b_j)\mathrm{d}x_i$来求出参数 a 和 b 的估计值。当然，在实际过程中，前式依然很难求解，更常用的是将求取期望和极大似然的步骤分开进行，具体如下：

1. E 步：根据前一轮得到的 a_j 和 b_j 的估计值计算出权重$p(x_i|a_j,b_j)$，并根据权重求出考虑隐变量情况下的完全数据对数似然的期望。
2. M 步：根据求出的对数似然的期望的表达式，求出使得该表达式取极大值时的 a_j 和 b_j 作为下一轮的参数。
3. 重复 E 步和 M 步直到达到期望的精度。

值得注意的是，在实践中即使在知道参数的情况下，E 步中的那个反常积分依然很难求，这个时候更常用的做法是将连续隐变量 x_i 离散化，并在离散的情况下改写期望，即构建一个函数$Q(x_i)$，使得$Q(x_i=\theta_k)=\gamma_k$ $(0<\gamma_k<1)$，且$\sum\gamma_k=1$，并将 E 步中的反常积分改写为：

$$\sum_k[\ln p(y_{ij}|\theta_k;a_j,b_j)]\gamma_k$$

到这里，一切看上去都还不错，我们的方法终于可以实施了……吗？

啊哈！聪明的小伙伴已经猜到了！我们之前选择权重的方式不总是合理的！

让我们回到 EM 算法中，这个方法希望通过使用完全数据对数似然的期望作为实际数据对数似然的下界，通过不断迭代，最终使得完全数据对数似然的期望的极大值点逼近实际数据似然的极大值点，这也就要求了我们的完全数据对数似然的期望是实际数据对数似然的紧下界，用人话来说，就是保证在完全数据对数似然的期望中至少有一点使得琴生不等式的等号成立。

现在，让我们回到一开始，我们逼近的目标是：

$$\ln L(a,b)=\sum_{i=1}^{n}\sum_{j=1}^{m}\ln\int_{-\infty}^{+\infty}p(y_{ij},x_i|a_j,b_j)\mathrm{d}x_i$$

首先，将隐变量 x_i 离散化（事实上，完全可以在假设的那一步直接设定隐变量是一个离散型随机变量，这样就省去了后面的诸多麻烦）：

$$\ln L(a,b)=\sum_{i=1}^{n}\sum_{j=1}^{m}\ln\int_{-\infty}^{+\infty}p(y_{ij},x_i|a_j,b_j)\mathrm{d}x_i\approx\sum_{i=1}^{n}\sum_{j=1}^{m}\ln\sum_{x_i}P(y_{ij},x_{ik}|a_j,b_j)$$

然后，引入权重$Q(x_{ik})$，$Q(x_{ik})$是一个分布列，$\sum_{x_i}Q(x_{ik})=1$，将对数似然函数重写为：

$$\ln L(a,b)=\sum_{i=1}^{n}\sum_{j=1}^{m}\ln\sum_{x_i}P(y_{ij},x_{ik}|a_j,b_j)=\sum_{i=1}^{n}\sum_{j=1}^{m}\ln\sum_{x_i}\frac{P(y_{ij},x_{ik}|a_j,b_j)}{Q(x_{ik})}Q(x_{ik})$$

显然，根据琴生不等式，有：

$$\ln L(a,b)=\sum_{i=1}^{n}\sum_{j=1}^{m}\ln\sum_{x_i}\frac{P(y_{ij},x_{ik}|a_j,b_j)}{Q(x_{ik})}Q(x_{ik})\geq\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{x_i}\ln\left[\frac{P(y_{ij},x_{ik}|a_j,b_j)}{Q(x_{ik})}\right]Q(x_{ik})$$

且等号在$\frac{P(y_{ij},x_{ik}|a_j,b_j)}{Q(x_{ik})}$为常数时取得，此时不妨设$\frac{P(y_{ij},x_{ik}|a_j,b_j)}{Q(x_{ik})}=C$，那么有：

$$\sum_{x_i}P(y_{ij},x_{ik}|a_j,b_j)=C\sum_{x_i}Q(x_{ik})=C$$

又注意到$\sum_{x_i}P(y_{ij},x_{ik}|a_j,b_j)=P(y_{ij}|a_j,b_j)$，此时可求得权重$Q(x_{ik})$为：

$$Q(x_{ik})=\frac{P(y_{ij},x_{ik}|a_j,b_j)}{C}=\frac{P(y_{ij},x_{ik}|a_j,b_j)}{\sum_{x_i}P(y_{ij},x_{ik}|a_j,b_j)}=\frac{P(y_{ij},x_{ik}|a_j,b_j)}{P(y_{ij}|a_j,b_j)}=P(x_{ik}|y_{ij};a_j,b_j)$$

又注意到，在我们假定了隐变量初始分布的情况下（不管怎样，总要有一个初始分布的），$P(x_{ik}|y_{ij};a_j,b_j)$这个东西实际上等价于在知晓了y_{ij},a_j,b_j的情况下，x_i的后验分布，即：

$$P(x_{ik}|y_{ij};a_j,b_j)\propto P(y_{ij}|x_{ik};a_j,b_j)P(x_{ik};a_j,b_j)$$

注意到个体知识水平这个隐变量与题目参数独立，且根据题设，$P(y_{ij}|x_{ik};a_j,b_j)$可以很轻松求出，故可以借助$P(y_{ij}|x_{ik};a_j,b_j)P(x_{ik};a_j,b_j)$这个“核”求出未归一化的权重，而后借助归一化的方式得到实际权重。到此为止，我们就可以借助 EM 方法求取题目参数的估计值啦。而在求出题目参数之后，隐变量也可以借助普通的极大似然进行估计了。

[一点也不]有趣？的思考题：

1. 本文介绍了单自变量 logistic 回归模型，请读者自行尝试推导多自变量情况下的 logistic 回归模型的似然函数。
2. 根据文中介绍的广义线性模型的思路，尝试推导 Possion 回归的似然函数。
3. 不难发现，本文最后的升级版例子实际上是 IRT（项目反应理论）中的 2-PL 模型（2 参数 logistic 模型），尝试在本文介绍的基础上完成对该模型的参数估计。[提示：通常情况下，可以假设对所有的个体而言，知识水平的先验分布是标准正态分布（本质上隐变量的单位和尺度是未定的，所以为了计算方便通常都直接按照标准尺度假设），离散化时，可以考虑在-4 至 4 之间均匀取点对分布进行离散化]
4. 事实上，EM 方法是一种应用场景相当广泛的方法，除 IRT 模型外，另一个需要使用 EM 方法的模型是高斯混合模型（Gaussian Mixture Model, GMM）该模型假设数据是从 K 个不同的正态分布（高斯分布）中抽取后以一定的比例混合而成。高斯混合模型具有如下概率密度：

$$f(x|\theta) = \sum_{k=1}^{K} \alpha_k \varphi(x|\theta_k)$$

其中α_k代表样本来自第 k 个正态总体的概率，$\sum_{k=1}^{K} \alpha_k = 1$，$\theta_k$表示第 k 个正态分布的参数向量，$\boldsymbol{\theta}_k = (\mu_k, \sigma_k)$，$\varphi$是正态分布概率密度函数，$\varphi(x|\theta_k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp - \frac{(x-\mu_k)^2}{2\sigma_k^2}$。

a). 假定数据集是由 3 个**均值不同**，**标准差不同**的一维正态分布混合而成，试求这三个正态分布的参数。

b). 假定数据集是由 3 个**均值向量不同**，**协方差矩阵相同**的二维正态分布混合而成，试求这三个二维正态分布的参数。

[提示：隐变量是样本来自哪个正态分布]

Ps. 对方法比较熟悉的小伙伴不难发现，实际上这道题描述的模型就是传说中的潜剖面分析（Latent profile analysis）模型。没想到吧，LPA 也是用 EM 法估计系数的.jpg