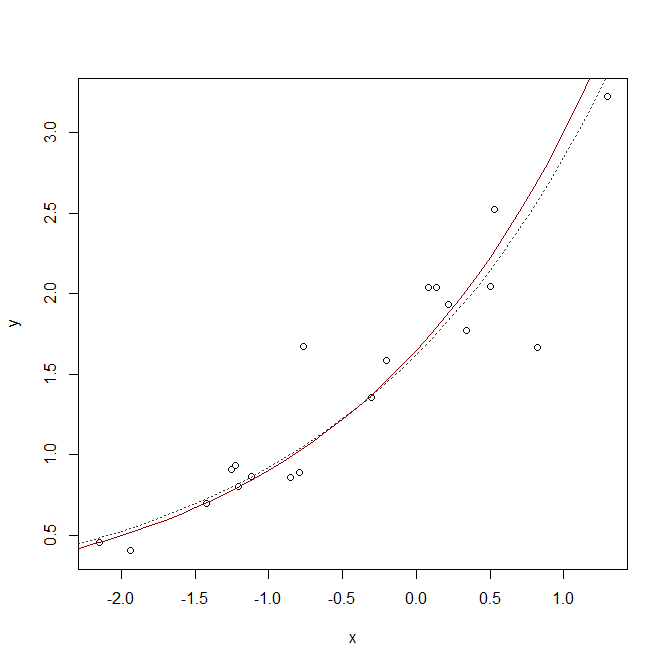
各位旅客朋友，您好，欢迎您乘坐椒盐海豹号列车，本次列车由线性回归开往IRT模型，请您上车后按照车票上指定的车厢和座位号对号入座。本次列车始发站为线性回归，途径变换后可化为线性回归的非线性回归、Logistic回归、广义线性模型、隐变量与EM方法，最终到达终点站IRT模型。

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）。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，对两边取对数，容易有，如果令，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，结果和生成数据所用的模型（）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。这种思路解决了绝大多数情况下的问题，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yi = 1，否则yi = 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，对于线性回归模型而言，当我们已知时，y服从均值为，标准差为σ的正态分布，此时容易有的可能性（也可以近似把它理解成“概率”）为：

那么，如果我们目前手头有一个简单随机样本, , …, ，并且我们假定x与y之间存在有的关系，那么很容易可以计算出当x为x1, x2, …, xn时，对应的y分别为y1, y2, …, yn的可能性为：

容易发现，在给定样本的情况下，上式是一个关于和的函数。进一步，求出当上式取最大值时的，这样的就可以作为参数的**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：

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，最大化等价于最小化，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，尺度参数为的对数正态分布，即对模型而言，当已知时，有y的概率密度为：  进而可以写出此时的对数似然函数为：  容易发现，此时最大化等价于最小化，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。 |

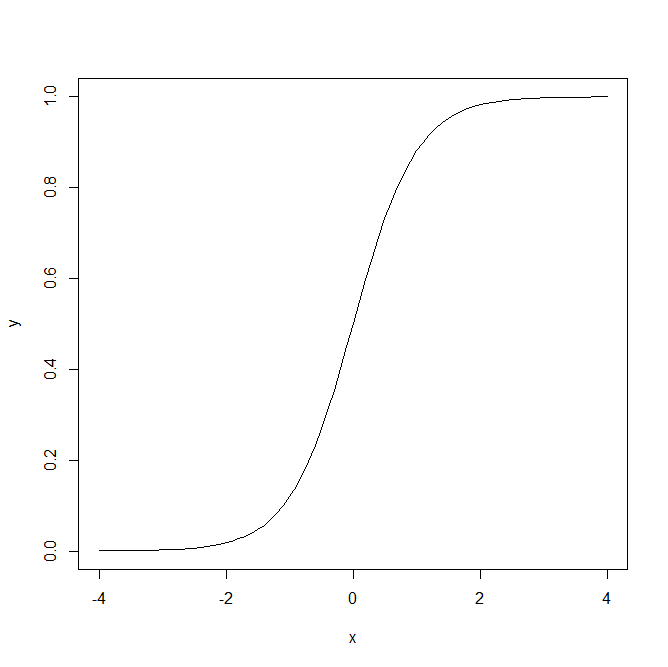
3. logistic回归

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：当已知时，以概率等于1，以概率等于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（）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：

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，假定个体的决策标准为ci，即只有当处理后的信号t’的强度大于tc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：  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θi，即时，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：  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：当已知时，以概率等于1，以概率等于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：

|  |
| --- |
| \*特别地，注意到当我们假设yi以概率等于1，以等于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：  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：  当我们令x1与x0相差一个单位（对于连续变量）或x1代表被试具有某属性x0代表被试不具有某属性（对于二分变量）时，容易有。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都是不完整的数据，中间遗漏了一个变量zi。换言之，实际的数据集是，但由于种种原因，我们只收集到了，但我们大概知晓下面两个信息：

1. 当时变量的分布
2. 当时变量的分布

特别地，我们假定当已知时，x与y条件独立，即。

注意到，此时的边际分布为：