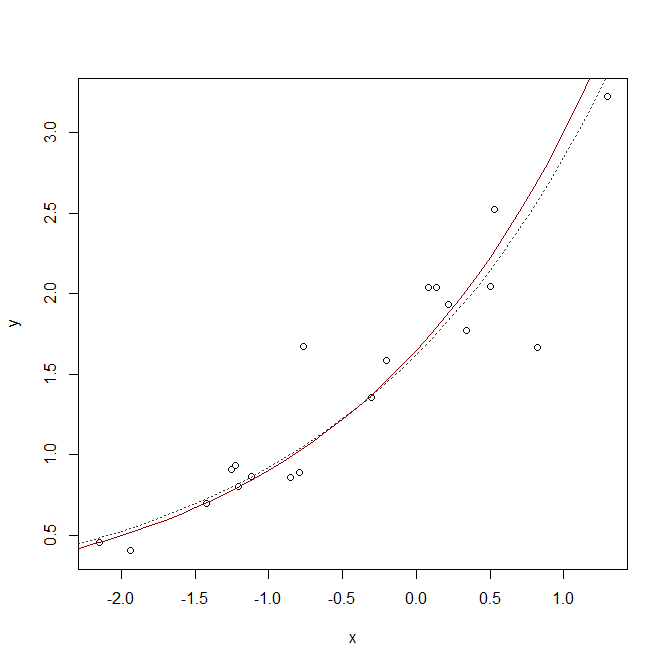
人不能，至少不应该在一天之内受到两次暴击.jpg【所以隔了几天更新？？？

Topic: 关于Logistic模型

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. 重回线性模型

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：

如果样本里的每一对x和y都是从具有这个特征的总体中抽取出的（注意虽然通常情况下我们习惯性假定(x, y)服从这是一个二维正态分布，但事实上回归模型并不要求自变量x也服从正态分布）。如果我们把从总体中抽取出的每一对样本记作, , …, 。那么，这个样本是从参数为这样的总体中抽出的可能性为：

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总体参数”。