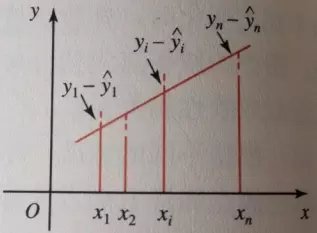
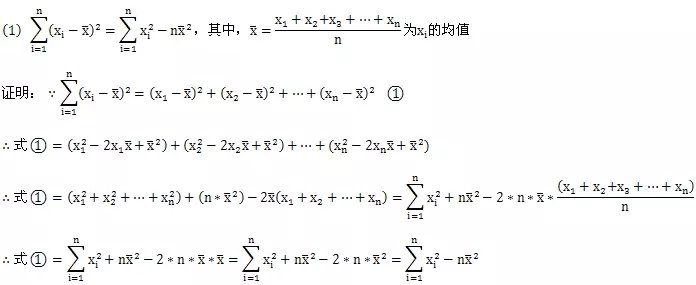
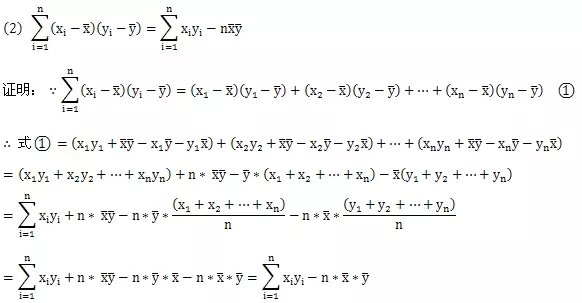
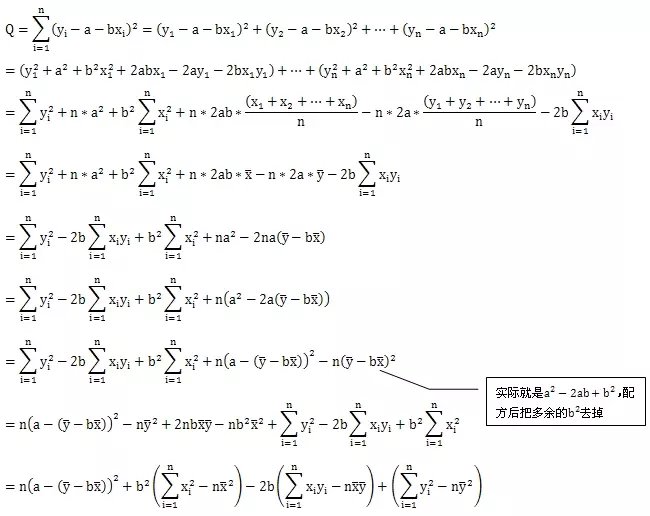
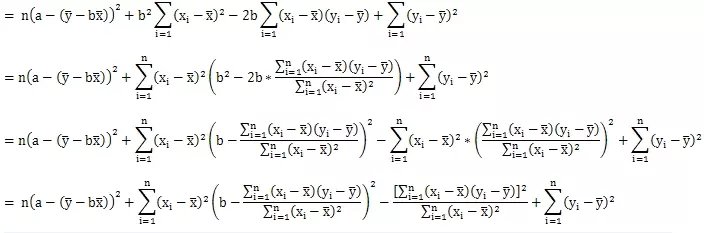
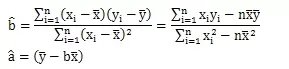
 在数据的统计分析中，数据之间即变量x与Y之间的相关性研究非常重要，通过在直角坐标系中做散点图的方式我们会发现很多统计数据近似一条直线，它们之间或者**正相关**或者**负相关**。虽然这些数据是离散的，不是连续的，我们无法得到一个确定的描述这种相关性的函数方程，但既然在直角坐标系中数据分布接近一条直线，那么我们就可以通过画直线的方式得到一个近似的描述这种关系的直线方程。当然，从前面的描述中不难看出，所有数据都分布在一条直线附近，因此这样的直线可以画出很多条，而我们希望找出其中的一条，能够最好地反映变量之间的关系。换言之，我们要找出一条直线，使这条直线“最贴近”已知的数据点，设此直线方程为：  
https://img-blog.csdn.net/20170221085349487?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvTWFyc0pvaG4=/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center  
  
这里的https://img-blog.csdn.net/20170221085442700?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvTWFyc0pvaG4=/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center是为了区分Y的实际值y（这里的实际值就是统计数据的真实值，我们称之为**观察值**），当x取值https://img-blog.csdn.net/20170221085510732?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvTWFyc0pvaG4=/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center(i=1，2，3……n)时，Y的观察值为https://img-blog.csdn.net/20170221085537769?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvTWFyc0pvaG4=/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center，近似值为https://img-blog.csdn.net/20170221085610175?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvTWFyc0pvaG4=/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center（或者说对应https://img-blog.csdn.net/20170221085649879?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvTWFyc0pvaG4=/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center的纵坐标是https://img-blog.csdn.net/20170221085714051?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvTWFyc0pvaG4=/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center）。  
  
其中https://img-blog.csdn.net/20170221085745205?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvTWFyc0pvaG4=/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center式叫做Y对x的回归直线方程，b叫做回归系数。要想确定回归直线方程https://img-blog.csdn.net/20170221085757565?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvTWFyc0pvaG4=/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center，我们只需确定a与回归系数b即可。  
  
 设x，Y的一组观察值为：  
https://img-blog.csdn.net/20170221085824925?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvTWFyc0pvaG4=/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center   i = 1，2，3……n  
  
其回归直线方程为：  
https://img-blog.csdn.net/20170221085853395?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvTWFyc0pvaG4=/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center  
  
当x取值https://img-blog.csdn.net/20170221090008959?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvTWFyc0pvaG4=/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center(i=1，2，3……n)时，Y的观察值为https://img-blog.csdn.net/20170221090030350?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvTWFyc0pvaG4=/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center，差https://img-blog.csdn.net/20170221090046444?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvTWFyc0pvaG4=/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center刻画了实际观察值https://img-blog.csdn.net/20170221090104772?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvTWFyc0pvaG4=/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center与回归直线上相应点纵坐标之间的偏离程度，见下图：  
  
  
 实际上我们希望这n个离差构成的总离差越小越好，只有如此才能使直线最贴近已知点。换句话说，我们求回归直线方程的过程其实就是求离差最小值的过程。  
  
一个很自然的想法是把各个离差加起来作为总离差。可是，由于离差有正有负，直接相加会互相抵消，如此就无法反映这些数据的贴近程度，即这个总离差不能用n个离差之和来表示，见下图：  
https://img-blog.csdn.net/20170221090224287?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvTWFyc0pvaG4=/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center  
   
一般做法是我们用离差的平方和，即：  
https://img-blog.csdn.net/20170221090246087?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvTWFyc0pvaG4=/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center  
  
作为总离差 ，并使之达到最小。这样回归直线就是所有直线中Q取最小值的那一条。由于平方又叫二乘方，所以这种使“离差平方和为最小”的方法，叫做**最小二乘法**。  
用最小二乘法求回归直线方程中的a、b的公式如下：  
https://img-blog.csdn.net/20170221090308850?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvTWFyc0pvaG4=/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center  




  
其中，https://img-blog.csdn.net/20170221090415229?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvTWFyc0pvaG4=/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center、https://img-blog.csdn.net/20170221090426433?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvTWFyc0pvaG4=/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center为https://img-blog.csdn.net/20170221090437647?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvTWFyc0pvaG4=/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center和https://img-blog.csdn.net/20170221090447089?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvTWFyc0pvaG4=/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center的均值，a、b的上方加“︿”表示是由观察值按最小二乘法求得的估计值，a、b求出后，回归直线方程也就建立起来了。  
  
当然，我们肯定不能满足于直接得到公式，我们只有理解这个公式怎么来的才能记住它，用好它，因此给出上面两个公式的推导过程更加重要。在给出上述公式的推导过程之前，我们先给出推导过程中用到的两个关键变形公式的推导过程。首先是第一个公式：  
 

接着是第二个公式：  
  
   
 基本变形公式准备完毕，我们可以开始最小二乘法求回归直线方程公式的推导了：  
  
  
   
 至此，公式变形部分结束，从最终式子我们可以看到后两项  
https://img-blog.csdn.net/20170221092002467?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvTWFyc0pvaG4=/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center  
  
与a、b无关，属于常数项，我们只需  
https://img-blog.csdn.net/20170221092039017?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvTWFyc0pvaG4=/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center  
  
即可得到最小的Q值，因此：  
  
  
至此，公式推导完毕。

最小二乘法求回归直线方程可用于所有数据分布近似直线的数据统计、分析问题，其用程序实现非常简便，属于基础统计分析算法，必须能够熟练掌握应用。