从forward stepwise regression到least angle regression

本文试图理清从逐步回归到最小角回归的发展脉络。并且给出一种更为广义的逐步回归的过程。即从算法的角度，来讨论线性回归模型参数估计的问题。然而笔者智力有限，无法深入理解这些算法的本质。因此，本文在结束时，笔者发现所写的东西都只是肤浅皮毛而已。需要学习的东西还是太多了。以下是本文正文，欢迎拍砖：

在前两篇博文中，已经给出了QR分解和SVD分解在求解线性模型回归系数时候的作用。这是两种基于矩阵运算的求解最小二乘估计的算法。它们将矩阵分解为更容易求逆的矩阵，以简化运算。

这里将要给出的算法，更符合“学习”的逻辑。尤其是最小角回归，其变量被选入和被剔除的求解过程，符合归纳逻辑的正反馈和负反馈，因此这种方法被纳入机器学习的算法是有其逻辑基础的。

在所有正式的叙述之前，要明确，我们已经将输入和输出都做了标准化处理。每个变量目前的均值均为0，标准差为1。即我们不考虑截距项。

逐步回归（Forward stepwise regression）

首先，还是来问一个问题：

“为什么我们需要逐步回归？？？”

当变量数多于样本数时，更一般的，在高维的情况下，自由度不足以估计所有的回归系数。因此需要进行变量选择。对于变量选择，最直观的想法是把所有子变量集的组合都列出来，然后每一个集合都进行最小二乘回归。得到若干个模型，然后根据某种准则从中选出最好的。这种普查型的择优选择，显然计算量会很大。因此我们需要寻找一条路径，在这条路径上进行优选，以保证最终得到的最优模型和普查所有模型下选出的最优模型差不多。

对启发式优化算法（比如模拟退火，遗传算法等）有所了解的同学，看到这里时一定会觉得很熟悉。没错，逐步回归其实就是一种最优化算法。有一种观点认为“最优化算法在统计学中占据核心地位”，以我浅薄的见解，对此我是十分赞同的。我们来大致描述该forward stepwise算法：

算法的目标是最小化残差平方和。

第一步，模型所有变量的回归系数都是0，我们需要找出一个变量，估计其回归系数，使得这个模型最能够减小当前的残差平方和，即最能改进模型的拟合情况。

第二步，从剩余变量中选出一个，使得加入这个变量之后的活跃变量集合能够使得模型拟合进步最大。用所有活跃集中的变量进行回归。

重复第二步，直到符合某种准则。

请注意，对于stepwise而言，每一步的关键是找出最能改进拟合情况的变量，然后对所有入选的变量，即被激活的变量集合重新做回归。也就是说每一步都做了多元回归。

逐渐回归（forward stagewise）

简单描述一下这个算法，在开始的时候，所有输入变量的系数都是0，即都处于睡眠状态，那么当前的残差就是输出*y*。

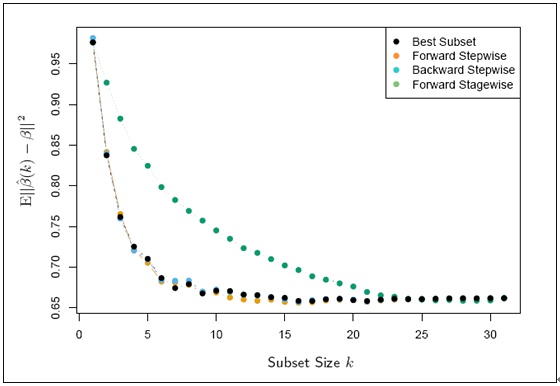
算法开始，找出和当前残差相关系数最高的变量，激活它，用当前的残差和其做回归，估计出回归系数。计算当前残差。

然后再找出所有变量中和当前残差相关系数最高的，激活它，用当前残差和其做回归，计算处回归系数，计算当前残差。注意这里在计算时，之前入选的变量的系数并没有变化。

然后在找出所有变量中和当前残差相关系数最高的（有可能是已经入选的变量），激活它，再用当前残差和其做回归，计算系数，计算当前残差。

重复上述过程直到达到某种准则。对于观测样本量大于变量数的情况，最终的结果就是普通的最小二乘估计。

从算法的描述我们可以看出，逐渐回归每一步并非充分地估足了回归系数。所以这种算法又被成为“慢拟合”（slow fitting）



上图可以看出逐渐回归算法的收敛速度要慢得多。

最小角回归（least angle regression）

在陈述对Efron及其LAR算法的崇敬之前，先来描述这种算法。我尽量试着用自然语言说清楚这个算法，不使用数学公式。

stepwise算法每一步激活一个变量，并且要对所有的活跃变量集合重新做回归来更新其系数。步子跨得比较大。而stagewise算法每一步激活一个变量，只更新该变量的系数。步子跨的又有点小。其减少了当前残差和当前被激活变量的相关性，但是经过重复类似的步骤之后，，以前被选入激活的变量和当前残差的相关性有可能又变成最小的而又被入选。这正是因为在每一步，只有该步入选的变量的系数进行了修改，而其他已入选变量系数没有变化。这样做每一步的步子确实跨得太小，收敛速度太慢。

而LAR算法正是对这种问题的一个修正。它每一步所走的路径，会保持当前的残差和所有入选活跃集的变量的相关性都相同。即同时修正所有入选变量系数，使得这些变量和当前残差的相关性会同步减少。

更多

上述的这些算法，可以用一个更通用的名称来概括——path algorithm。这些算法和启发式优化算法具有很多相似的地方。启发式优化算法在统计学中所占地位是非常重要的，也越来越受到重视。