基于多元线性回归对棉花产量影响要素的研究

摘要

棉花是世界上最主要的农作物之一，是国民经济的命脉，但影响棉花单产的要素有许多，合理的资金投入不仅可以增大面产量，还能降低生产成本。

本文采用多元线性回归探究所给要素对棉花单产的作用力，进一步采用逐步回归的方式筛选出重要的投入要素。

关键词：回归分析，逐步回归

模型的建立

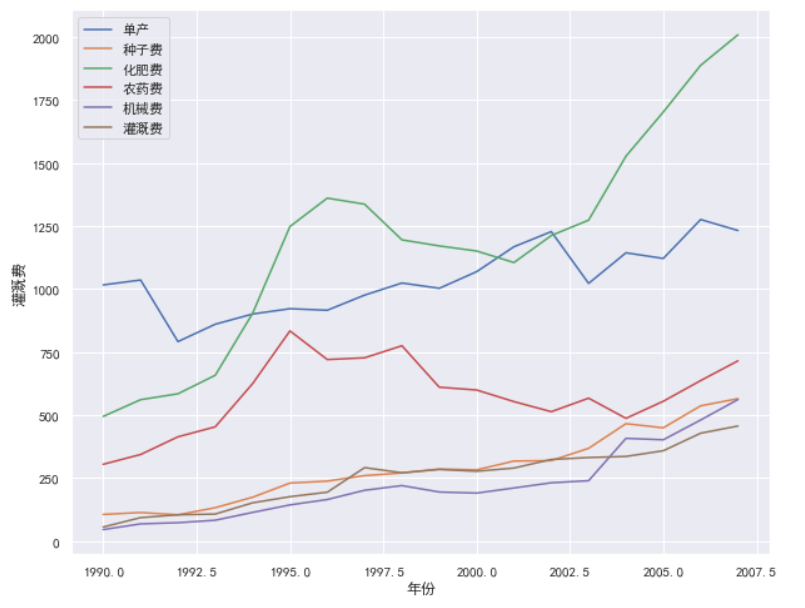
多元线性回归模型的基本形式如下：

单产 = β0+β1种子费用+β2花费费+β3农药费+β4机械费+β5灌溉费

其中单产单位为kg/公顷，其余费用单位均为元/公顷。针对上述变量进行描述性统计，

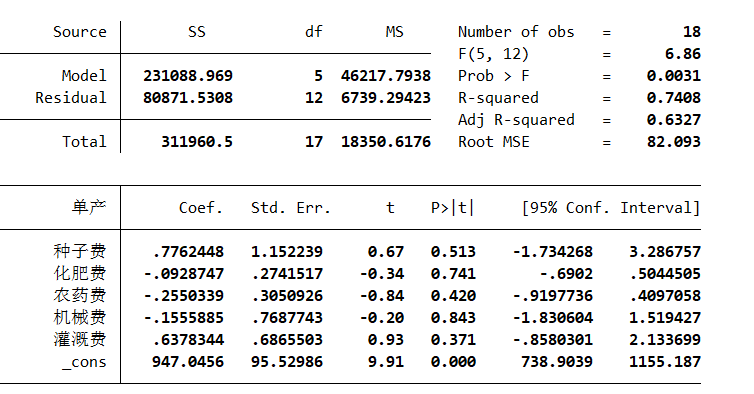


从表中min一栏可以看出，化肥费与农药费是投入比例最多的两种费用，而机械费与灌溉费相对较少。



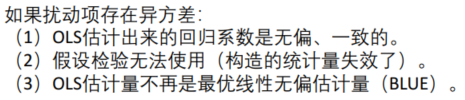
从上图可以看出，1990-2007年间花费费用的投入大幅上升，其余费用投入相对平稳。

多元线性回归模型求解

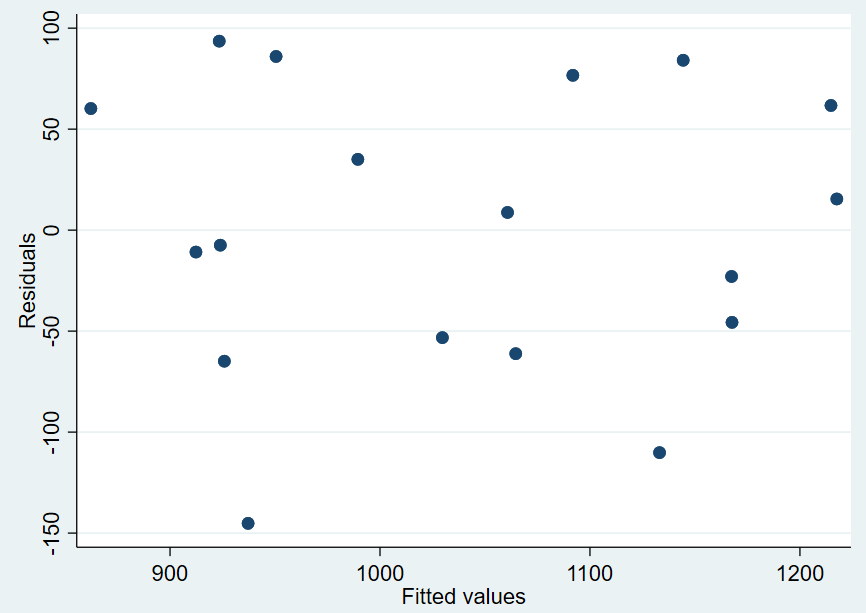
利用stata软件对上述数据进行多元线性回归，回归结果如下：

Prob > F = 0.0031，说明在回归系数总体上在95%的置信水平上是显著的，但进一步发现所有指标单独来看都是不显著的。

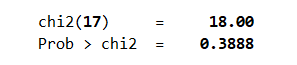
异方差检验：



检验结果如下：



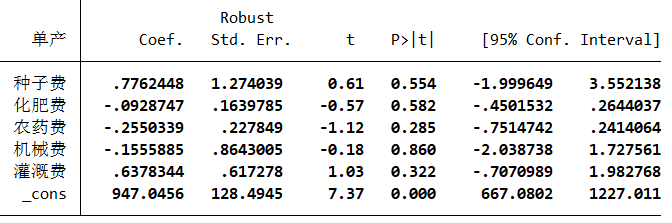
可以看到，残差与拟合值的分布毫无规律，这也验证了上述指标的不显著性是正确的，

怀特检验：

Prob > chi2 = 0.3888，说明在95%的置信水平下，模型不存在异方差，与上述得出了相反的结论（离谱）。

OSL+稳健的标准误

为了增加显著的变量，仍进行OLS回归，但使用稳健的标准误进行修正，

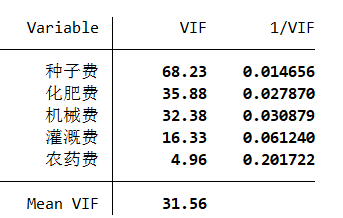


对比普通OLS回归结果，各变量的显著性没有提高反而下降（再次离谱）。

多重共线性检验

进一步考虑模型存在多重共线性的可能性。

方差膨胀因子(VIF)是关于自变量与其余自变量回归得到的拟合优度的函数，拟合优度越大说明某个自变量与其余自变量之间的相关性越大，认为模型存在严重的多重共线性。在回归结束后利用stata的estat vif命令可以计算各个变量与其余变量间的VIF。结果如下：

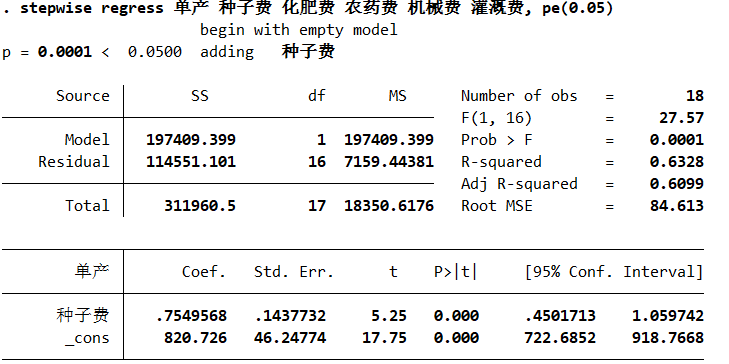


通常认为VIF>10说明回归方程存在严重的多重共线性，从结果中可以看到，除农药费以外的所有变量都没有通过检验。可以初步认为，农药费是影响棉花单产的决定性因素。

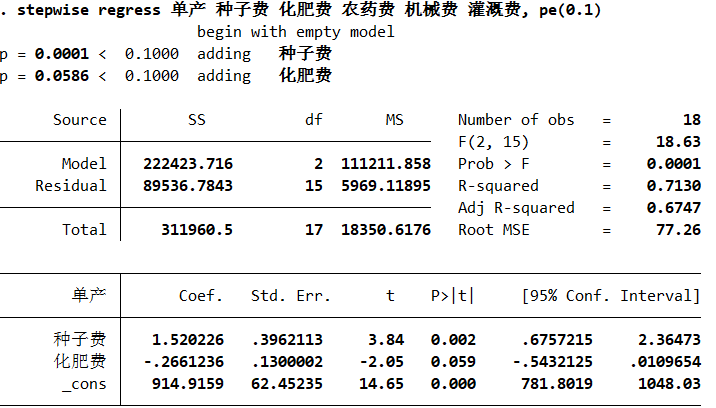
逐步回归分析

对于存在多重共线性的模型可以使用逐步回归分析来剔除不显著的变量。逐步回归包括向前逐步回归与向后逐步回归，其原理都是通过计算整体模型解释因变量的能力来筛选不显著的变量，向前逐步回归向模型中逐个增加变量，而向后逐步回归最初引入所有变量，之后尝试将某个变量从模型中剔除来提高模型的解释性，两种回归方式最终都保留高显著性且解释力强的变量作为最终变量。

向前：

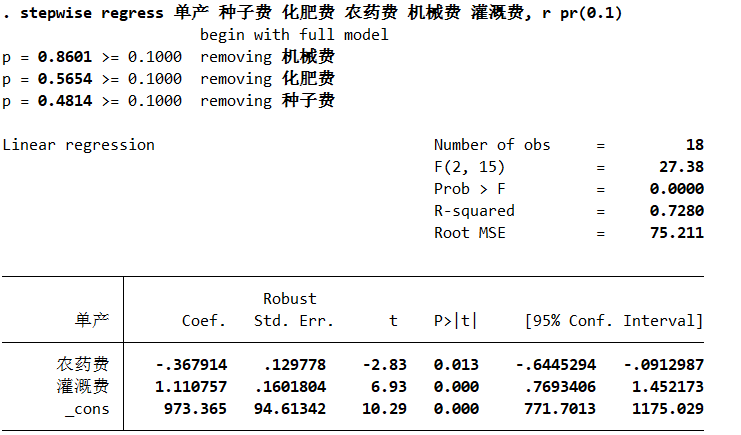


在设置的显著性为0.05的水平上进行向前逐步回归仅保留了种子费一个变量，尝试增加显著性水平来保留更多的变量。



在设置的显著性为0.1的水平上保留了种子费与化肥费两个变量作为最终变量。

向后：



在设置的显著性为0.1的水平上，向后逐步回归保留了农药费与灌溉费作为最终变量，可以看到与向前逐步回归分析的结果不同。