经常看到做变量的相关性分析，对于做线性分析而言，变量之间的相关性分析到底有什么用？

引用别人的博客：

最近把统计学读完才终于弄明白信用风险模型中所说的多重共线性到底是什么，为什么要求最终模型中一定不能存在多重共线性，本以为这些变量不是都对预测违约率有作用吗？那留着为什么不行呢?希望通过这篇文章让自己彻底搞明白

**什么是线性回归模型？**

先聊聊啥是线性回归模型，*线性回归模型=线性+回归模型*

**回归分析**

用一定的数学表达式讲变量之间的数量关系描述出来，进而确定一个或几个变量（自变量）的变化对另一个特定变量（因变量）的影响程度  
**解决什么问题**  
1最显然的可以利用所求的关系式可估计或预测因变量的取值，并给出预测的可靠程度，比如“明天下雨的概率”  
2附带可以知道在影响因变量的诸多变量中哪些变量的影响是显著的，哪些是不显著的，比如遇到领导问“最近的客群很奇怪，帮我看看到底这些人有什么特征？”“你就知道线性回归分析是可以用来解决这个问题的”

**回归模型**

就是描述因变量y如何依赖于一个或多个自变量和误差项e的方程

**线性回归**

当然指的是因变量和自变量之间的关系是线性的

**一元/多元线性回归**

区别是自变量的个数

举个多元线性回归的栗子 y=b0+b1x1+b2x2+...+bkxk，其中b1,b2称为偏回归系数，b1表示当x2,...xk不变时，x1每变动一个单位因变量y的平均变动量，其余偏回归系数的含义也是类似的。

**多重共线性问题**

回归模型中存在两个或两个以上的自变量彼此相关

**有啥影响**

1变量之间高度相关，可能使回归的结果混乱，甚至把分析引入歧途。  
2对参数估计值的正负号产生影响，特别是估计系数的符号可能与预期的正相反，造成对回归系数的解释是危险的。  
比如：违约率应该和贷款余额是正相关的，但由于有其他因素的影响最终模型中贷款余额的系数为负，得到“贷款余额越大违约率越低”的危险解释。  
**可见：在建立线下回归模型时，不要把所有自变量都放进来除非有必要，因为他们带来问题比解决的问题可能更多！**

**如何判别多重共线性**

1如上文所说，回归系数的正负号与预期的相反  
2模型的线性关系检验（F检验）显著，但几乎所有回归系数bi的t检验却不显著（后面文章会详细解释线性关系检验和t检验）  
3容忍度和方差扩大因子VIF（variance inflation factor）本文主要解释的方法  
VIF=1/（1-Ri平方）  
某个自变量的容忍度=1减去该自变量为因变量而其他k-1个自变量为预测变量时所得到的线性回归模型的判定系数=1-Ri平方

其中，判定系数Ri平方=SSR/SST=回归平方和/总平方和，测度了回归方程对观测数据的拟合程度，若所有估计值与观测值都完全一致，Ri平方=1

那么容忍度越小，第i个自变量被其他k-1个自变量可拟合的程度就越高，代表多重共线性就越严重

而VIF=容忍度的倒数，VIF越大，多重共线性就越严重，一般认为VIF大于10时，存在严重的多重共线性

**如何解决多重共线性问题**

1简单粗暴，把相关的自变量从模型中剔除  
2采用PCA等变量降维方法，原始变量一个不要全部采用PCA组合出来的变量，一次性解决多重共线性问题  
3在建模过程中采用逐步回归的方法逐个选择进入模型的变量，避免多重共线性问题发生  
**选择变量的方法**  
一个变量是否留在回归模型中的检验根据是，使残差平方和SSE显著减少，方法有向前选择，向后剔除，逐步回归，最优子集等  
以向前选择法为例说明过程：  
第一步：对k个自变量分别拟合因变量y的医院线性回归模型，共k个，找F统计量值最大的自变量xi，第一个引入模型  
第二步：在第一个变量的基础上，分别引入其他k-1个自变量，建立k-1个回归模型，变量组合分别为xi+x1，xi+xi-1，xi+xi+1，xi+xk，找到F统计量最大的模型，并把模型中的xj引入模型  
依次不停的增加自变量，直到增加自变量不能导致SSE显著增加为止。  
向后法与向前法相反，逐步回归就是将上述两种方法结合起来。

上述基本把多重共线性问题的来历和影响解释清楚了，至于文中提到的F检验和t检验到底是什么怎么做会在后续文章中一一mark下来

我会出一个关于信用风险建模系列文章，即是对自己学习的总结也是希望与大家分享交流共同进步

知其然更要知其所以然，加油！

结论：如果相关性太高，不要把全部变量扔进去，而是一个X一个X的扔进去跑benchmark



