**第一次：**

吐血的给你写一份数据使用说明，我下载下来已经要吐血了。。。。。

我不知道python咋回归，但是应该还是要整理成面板数据的，但是这里有一个问题，我下午要问老师的，你要是有空可以先想想。就是非平衡面板数据和平衡面板数据，我看了很多关于上市公司的研究最后都是把非平衡面板数据处理成平衡面板数据了，我不知道是不是一定要这样做。他的做法应该是把不是每个截面上都有数据的样本删掉了，然后缺失部分指标的样本也删了，最后使得每个截面上的个体是一样的。就像有的公司后来才上市，不是在2011-2018年整个跨度都有数据的，那这种样本删掉，然后对于部分指标缺失这种情况我有点疑惑，刚才大概看了下数据，发现政府补贴和绿色专利都会存在某个截面上为0的情况，那这种会不会造成影响还是应该把这种数据删掉，就感觉，哦哦对这个可以通过控制年份效应去实现。

我尝试用vlookup通过股票代码去匹配了，简直吐血，然后我数据透视表用的不是很好，实在是弄不成面板。。。。。

压缩包里有一份制造业上市公司股票代码excel（8），初步样本是2458个，但后面处理后不知道还会剩多少个样本，我觉得按之前的文献来看，最后剩下个800-1300个公司都是合理的。然后这里面除了**绿色专利（4）**其他的数据都是我从国泰安数据库下载的，在下载之前已经筛选了字段，只挑了对这个研究直接或间接有用的字段，然后每个压缩包里的txt文件是对各字段的含义、赋值、单位的说明。（压缩包4是单独下载的，不是从数据库，所以里面包含了全体上市公司，要筛选一下，筛选出A股制造业上市公司，根据文件9）

压缩包（1）董监高个人特征文件：年龄、性别、受教育程度（学历）

这个麻烦的是，这个要先把它变成截面数据，把一个时间一个公司的所有高管的年龄啊什么平均下，你看那个变量定义表就行。这个表下载下来的时候已经筛选过了，这里面都是高管。但是压缩包2无法筛选，就是还包含高管以外的人。所以最好压缩包2按照1里面的人来匹配。

压缩包（2）董监高任职情况表：任期

压缩包（3）公司基本情况文件：成立日期

**压缩包（4）绿色专利数据：绿色专利授权量（获得量）这份数据里面的公司不仅仅是制造业，所以要根据文件（9）去筛选一下。**

压缩包（5）政府补助：看每年合计数据那一行就ok

压缩包（6）利润表：营业收入、净利润。这份文件因为涉及到营业收入增长率的计算，所以时间跨度是2010-2018而不是2011-2018，研究设计中有计算公式，营业收入增长率=（本年营业收入-上年营业收入）/上年营业收入，然后注意下财务报表都是年末12.31公布的。

压缩包（7）资产负债表：总资产

压缩包（8）中国上市公司股权性质文件：股权性质及编码。与研究设计中的不同，要转换下

压缩包（10）财务指标文件：ROA ROE

压缩包（11）研发投入情况表：

表12绿色专利授权量数据(2001-2018)：**绿色专利授权量（获得量）这份数据里面的公司不仅仅是制造业，所以要根据文件（9）去筛选一下**

**第二次修改：**

来孩子我告诉你下，改哪儿。我错了我错了，那个应该用A表合并报表的数据，然后下面第一个文件是你第一次发给我的，我发现合并报表包含更多信息，然后很多因为缺失值被删掉的数据用A表的数据的话就不用被删掉。我看了下，就资产负债表和利润表涉及到A表B表的问题，我再把要用的发你下筛选好的，然后利润表两个量（净利润、营业收入）都用12-31日的作为当年的，不需要加总。跟之前还是一样，删掉12-1-1日之后上市的公司样本，表中总资产和**营业收入增长率**、净利润三列数据要改，最后再删掉有缺失值的案例行。

然后第二个问题是Aedu，就按我前面说的改：看那个字段说明，1-5其实都是按照一个层次顺序上去的，但是6我觉得就不是很能衔接在一起，所以我想你把原来高管特征里学历所有为6的删掉，然后再求平均。

**第三次修改：**

用这个表 还有do文件回归了一下，你可以试一下看的更清晰一点， 然后就是模型（2）做出来 第2、3个调节变量是显著的 然后模型（3）做出来调节变量2、3、4是显著的（0.05显著性水平上，但在0.01不显著） ，用模型（3）这个结果是可以的，但是一般来说模型(2)更规范一点，所以想后面调整变量使得模型（2）显著。其实最后可以删掉1-2个调节变量，当然保留也没有问题。就是现在主要的问题有两个：1.主效应不显著，就是政府补贴对绿色创新的影响，即模型（1），然后我觉得这个跟我绿色创新选的指标有点关系，绿色专利那个数据感觉变差有点小，不是很有趋势，然后这里就是想换个指标试一试结果。 2发现控制变量选的不是很好，我再去查阅一下这方面的相关文献，换或者增加1-2个控制变量。 （我上面说的这些模型几模型几你看下do文档哈）

来，我来说明一下，新增的3张数据表在压缩包里。然后你尽量在“1原始表”的基础上去弄，这是你第一次没有删减过缺失值还有12-1-1后上市的样本的，只是我取了对数并且为了方便改了变量名（用这张表的原因是，我后面需要试有不同变量的模型，我担心含不同变量进行缺失值删除情况不一样，所以让你用最原始的表来弄，对了你匹配好数据后这张原始的表也发我一下我备份），然后需要匹配数据的列我都在excel标明并且备注了，对了因为我这边表不全，所以那个Soe还是没有转换成0和1的，你那边要是有转换过的直接弄上去，要是没了麻烦再转换一下。你要做的事情是1把几列空白的根据压缩包里数据匹配上去 2（因为前面的检验显示Soe不是很显著，所以我想试试控制变量有Soe和无Soe的差别），就是你删除缺失值的时候分两种情况，一种不要考虑Soe,一种考虑Soe。 最后给我三张表，一张匹配过的原始表，一张考虑Soe的最终面板数据表，一张未考虑Soe的最终面板数据的表。