**模型的修改：**

**模型求解中的问题：**

在模型最初的构建时，我们希望通过一个矩阵A\_ij来表示县i对县j的影响，进而完善part1中的问题求解。但是在part1中，我们使用的数据其实只有2010-2017年一共八年的数据，并且每年的数据我们真正利用的部分只有每个州毒品数量的总和，还有每个县当年对应的数量。所以，我们求解整个模型的时候，只能够采用抽象的方式考虑两个县之间的影响，虽然县i可能对县j的影响是方方面面的，但是我们统一采用aij来表示这种影响关系，也就是让aij作为所有因素的综合考量。

同时在part1的模型简化过程中，还存在一定的问题。比如说，我们必须得知道两年的数据，才能够求解出两个县这一年的影响强度。换句话说，我们需得到“未来”的数据，才能够知道某个县与县之间的影响强度。但是这在现实世界是完全不可能的，如果我们已知下一年每个县对应的毒品数量，那这个模型的建立也就失去了意义。但在part1里，我们其实不得不采用这种方式计算。如果细致探究县与县之间的影响关系，我们需要足够求解矩阵461\*461个（通过统计，在本题的求解中共有461个县）未知变量的方程组，也就是说，我们起码得有400多年的数据才能够求解（还要保证，这400多年中一个县对另一个县的毒品传播强度始终不变）。但是一旦我们拿到了详细的数据，我们就将对我们part1中的模型求解方式进行修改。具体修改将在下面详细叙述：

**模型的修改细节：**

Part1中，模型求解最大的问题就是我们需要“未来”的数据来评估传播强度。我们修改的过程将对这个重点关注。在part2中，我们被提供了美国社会经济数据，这些数据含有诸如全县的家庭数、家中有小于18岁孩子的家庭数等。目前存在几种假设来猜测为什么毒品数量达到了目前的水平，我们需要关注的人群有哪些，美国社会经济因素数据是否对毒品的传播有一定的影响。其实总结下来，我们只需要找到美国社会经济因素的数据对当年这个县的毒品数量的影响，且发现两个县之间这些社会经济因素的数据对两个县之间毒品传播有哪些影响，也算是完成了对这个模型的修改并求解，进一步通过我们构建的模型完成对未来毒品传播走向的估计，并及时提醒相关部门采取相关政策控制当地毒品的传播，同时利用模型来检测这个政策将会对毒品的传播有怎样的影响（将在part3的讨论中给出）。

我们继续采用a\_ij作为县i对县j的传播影响强度，同时我们认为每个县当年的毒品数量和该县当年的所有因素都是相关的（具体相关性的大小将通过因素对应的权重进行描述），同时，我们假设每个县关于这些因素的相关性都是一样的，也可以说是权重系数是一致的。

在这种前提假设下，我们不用去直接求解461\*461个未知变量，我们可以考虑下列带有k个（其中k为part2允许使用的因素树）未知量的线性函数。也就是下列表达式所示：

Yi = r1\*xi,1 + r2\*xi,2 + … + r0 + δ + ε

其中，Yi代表这一年县i的毒品数量，x1， x2等代表该县当年在某因素限制下符合条件的数量， r1， r2等分别代表毒品数量和x1等因素的权重系数，在一定层面代表着毒品数量和该因素的相关性，同时r0等都代表偏置。

具体各个因素求解的结果将放在附录的表格中供参考。因为我们希望通过part2中对模型的修改，能够得到每个县各个因素对该县本身的影响，同时可以通过两个县之间每个因素的关系来找到两个县之间毒品数量的联系。为此我们引入Yi,j表示县i和县j之间毒品数量的差。很显然我们有以下两个等式关系：

Yi,j = Yi – Yj

Yi,j = r1(xi,1 – xj,1) + r2(xi,2 – xj,2) + … +

也通过这种运算，将各种偏置项的影响消除。

根据part1的假设，我们只考虑源头（当年爆发点所在）和某个县的影响。也就是说，上式中的i和j将会有一个是源头，这里我们定i为源头进行下面的分析。Yi,j代表源头县i对县j的毒品数量的差。同时根据Part1的基本假设，我们认为两个县之间的相关性，为其中不为源头的县j两年的差值△Yj/Yi。但是part1我们是通过下一年的数据来完成两个县之间的相关性的求解，这里我们修改为

βi,j = Yi,j / Yi

= (r1(xi,1 – xj,1) + r2(xi,2 – xj,2) + … +)/Yi

通过这种计算，只要我们知道每个非源头的县j 和 源头县i各个因素对应的数值xi,1…，我们就能够通过当年的数据完成对所有县下一年毒品数量的检测。但是这一切的前提都是，我们完成对于r1等因素和毒品数量的相关性大小的计算，也就是权重系数的计算。下面讲述，我们权重系数的求解过程。

对于权重系数，我们利用

Yi = r1\*xi,1 + r2\*xi,2 + … + r0 + δ + ε

求解。因为我们含有2010-2016的人口普查结果，同时可以通过part1的表格得到每年对应的Yi值，提供了足够大的数据量进行选择。我们考虑到对于数百个未知量的求解只需要几百个方程即可，但是我们有着七年的数据，远远超过所需要的方程数。同时，考虑到社会经济因素的统计以及毒品数量的统计可能存在误差。

为了将这种个体数量的影响充分降低，我们采用机器学习中的梯度下降的方式对每一个参数进行修正。我们首先将每个参数赋值为0，如果通过梯度下降的思想对每个值进行修改和检验，最后完成了这些值的计算，模型修改成功。

**模型的应用：**

**未修改前：**

在part1中，我们通过简化模型并简化模型的计算，通过给定的较少的数量完成了对每个县的毒品数量的估计并取得了不错的结果。首先我们完成了对于每个州源头的寻找，分别是Kentucky州的JEFFERSON，Ohio州的HAMILTON, Pennsylvania州的PHILADELPHIA, Virginia州的FAIRFAX, West Virginia州的KANAWHA。

几个需要特别关注的县（意味着会在未来取代源头并成为新的爆发点）分别是：2010-2017年中 Ohio州的CUYAHOGA已经爆发，2019年是West Virginia的WOOD,2022年是Pennsylvania的ALLEGHENY,2025年是Kentucky州的KENTON。达到的药物阈值，分别是26715起，499起，9299起，2146起，分别达到了各自州的顶峰。其中需要尤其重视的地区是，Ohio州以及对应的CUYAHOGA县，该州的毒品数量远远超过其他各州顶峰的总和，且上涨速度非常恐怖。

**模型修改之后：**

模型修改之后，我们得出了每个因素对于毒品数量的影响。我们采取了两种估计方式（分别是广义最小二乘和极大似然估计）得到各个因素对毒品数量的影响是否显著。这里值得一提的是，由于多重贡献性（多个因素共同构成了另外一个因素，就比如该县的各种类型的家庭个数和该县总的家庭数非常接近）的影响，我们将这些因素直接抛除，在后续单独对每个由于多重贡献性被剔除的因素进行分析。

在两次检验中，我们各自得到了一组相关性尤其大的因素，同时挑选其中两次都被选出来的因素。这些因素中呈现正相关的有，该县中男性主人且有孩子无老婆的家庭数、未婚女性（孤寡、离婚）家庭数、祖父母养孩子年份较长家庭数、意大利和捷克等外来人口家庭数等因素都直接导致了毒品数量高水平。

由于多重贡献性被剔除的因素中，有着25岁以上人口、非社会福利人口、说其他地方语言的人口等因素，都直接导致毒品数量上升。

对于这些因素的考察，给了我们极大的信息量，同时也指示了我们，除了打压毒品等明面上的政策，同样有哪些潜在的社会经济因素需要考虑，将在part3中具体阐述并给出相应的政策建议。