**Modification of the model:**

**Problems in model solving:**

In the initial construction of the model, we hope to use a matrix A\_ij to represent the impact of county i on county j, so we can then solve the problem in part1. However, in part1, we used only eight years of data from 2010 to 2017, and the actual data we use is only total number of drugs in each state, and also the number of corresponding counties in each county. Therefore, when we solve the whole model, we can only consider the influence between the two counties in an abstract way. Although the influence of county i on county j is in all aspects, we use aij to express this integrated influence relationship as to make aij a comprehensive consideration of all factors.

At the same time, there are still some problems in the model simplification process of part1. For example, we must know the data for two years to be able to solve the impact intensity of the two counties this year. In other words, we need to get the “future” data to know the the impact between two counties. But this is completely impossible in the real world. If we know the number of drugs in each county in the next year, then the establishment of this model will lose its meaning. But in part1, we actually have to calculate this way. If you carefully explore the relationship between counties, we need to solve the matrix of 461\*461 (by statistics, there are 461 counties in the solution of this problem) unknown variables. That to say, we need at least 400 years of data can the problem be solved (and it should be guaranteed that the spread of drugs from one county to another in the past 400 years will remain the same). But once we have the detailed data, we will modify the way the model is solved in our part1. The specific modifications will be described in detail below:

**Modification details of the model:**

In Part 1, the biggest problem with model solving is that we need “future” data to assess propagation strength. Our revised process will focus on this. In part2, we were provided with US socio-economic data, such as the number of households in the county, and the number of households with less than 18 years of age at home. There are several hypotheses to guess why the number of drugs has reached the current level. What are the people we need to pay attention to? Does the US socio-economic factor data have a certain impact on the spread of drugs? In fact, we only need to find out the impact of the US socio-economic factors on the number of drugs in the county that year, and find out how the data of these socio-economic factors between the two counties affect the drug transmission between the two counties. It is completed to modify and solve this model, and further complete the estimation of the future drug transmission trend through the model we built, and promptly remind relevant departments to adopt relevant policies to control the spread of local drugs, and use the model to test how this policy will affect the spread of drugs affect (as given in the discussion of part3).

We continue to use a\_ij as the intensity of the influence of county i on county j. At the same time, we believe that the number of drugs in each county is related to all factors of the county (the specific correlation will be described by the weight of the corresponding factor). At the same time, we assume that each county has the same correlation about these factors, and it can be said that the weight coefficients are consistent.

 Under this premise, we don't have to solve 461\*461 unknown variables directly. We can consider the following linear functions with k unknown variables (where k is the number of factors allowed by part2). This is the following expression:

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

where Yi represents the number of drugs of county i , x1 represent the number of eligible conditions in the county under the constraints of a certain factor in the current year, r1 represent the weight coefficient of factors respectively. It represents the correlation between the number of drugs and this factor, while r0 and other variables represent offsets.

The results of the specific factors will be listed in the appendix for reference. Because we hope to modify the model in part2, we can get the influence of each county's factors on the county itself, and we can find the relationship between the two counties through the relationship of these factors between each of the two counties. So we use Yi, j to indicate the difference in the number of drugs between county i and county j. Obviously we have the following two equation relationships:

Yi,j = Yi – Yj

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

This operation is also used to eliminate the effects of various bias terms.

According to the assumption of part1, we only consider the impact of source (the point with most of drugs of the current year) and a county. In other words, we will have a source in either i and j in the above formula. Here we will determine i as the source for the following analysis. Yi,j represents the difference in the number of drugs in the county. At the same time, according to the basic assumption of Part1, we believe that the correlation between the two counties is the difference △Yj/Yi of the county j which is not the source. But in part1, we use the next year's data to complete the correlation between the two counties, here we modify the model into this:

βi,j = Yi,j / Yi

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

Through this calculation, as long as we know the value xi,1... corresponding to each factor of each non-source county j and source county i, we can complete the detection of the drug quantity in all counties of next year through the data of the current year. But the premise of all this is that we complete the calculation of the correlation between factors such as r1 and the number of drugs, that is, the calculation of the weight coefficient. The following describes the solution process of our weight coefficients.

For the weight coefficient, we use following to solve:

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

Because we have the 2010-2016 census results, we can get the corresponding Yi value every year through the form of part1, providing a large enough data to choose. We consider that only a few hundred equations are needed to solve hundreds of unknowns, but we have seven years of data, far exceeding the number of equations we need. At the same time, there may be errors in considering statistics on socio-economic factors and statistics on the amount of drugs.

In order to fully reduce the impact of this number of individuals, we use a gradient algorithm in machine learning to correct each parameter. We first assign each parameter to 0. If each value is modified and tested by the idea of gradient descent, the calculation of these values is completed and the model is successfully modified.

**Model application:**

**Before modification:**

In part1, we simplified the model and simplified the calculation of the model, and achieved an estimate of the number of drugs in each county with a given small number and achieved good results. First we completed the search for each source in every state, and they are JEFFERSON in Kentucky, HAMILTON in Ohio, PHILADELPHIA in Pennsylvania, FAIRFAX in Virginia, KANAWHA in West Virginia.

Several counties that require special attention (meaning that they will replace the source and become a new outbreak in the future) are: CUYAHOGA in Ohio State has already exploded in 2010-2017, WOOD of West Virginia in 2019, and ALLEGHENY of Pennsylvania in 2022, KENTON of Kentucky in 2025. The drug thresholds reached 26715, 499, 9299, and 2146, respectively, which reaching the peak of their respective states. Among the areas that need special attention, the state of Ohio and the corresponding county CUYAHOGA need more concerns, where the number of drugs far exceeds the sum of the peaks of other states, and the rate of increase is terrifying.\

**After modification of the model:**

After the model was revised, we came up with the impact of each factor on the number of drugs. We have adopted two estimation methods (generalized least squares and maximum likelihood estimates, respectively) to determine whether the effects of various factors on the number of drugs are significant. It is worth mentioning here that due to the multiple contributions (multiple factors together constitute another factor, such as the number of families in the county and the total number of households in the county are very close), we will throw away these factors, and each factor that is excluded due to multiple contributions is analyzed separately.

In each of the two tests, we each obtained a set of factors that were particularly relevant, and selected all factors that were selected twice. Among these factors, there are positive correlations between the number of families having male with a child and no wife in the county, the number of unmarried women (solitude, divorce), the number of families with long-term education under grandparents, and the number of migrants in Italy and the Czech Republic. All these factors have directly led to a high level of drugs.

Among the factors that have been excluded from multiple contributions, factors such as the population over 25 years old, population with non-social welfare, and the population speaking other local languages ​​have directly led to an increase in the number of drugs.

The investigation of these factors has given us a great amount of information. At the same time, it has also indicated to us that in addition to the policies on the bright side of suppressing drugs, there are also some potential socio-economic factors to be considered, which will be elaborated and given in part3 with the corresponding policy recommendations.