武漢大學
WUHAN UNIVERSLIY

科学软件网
SCIENCE SOFTWARE NETWORK

Stata研讨会

新冠疫情数据的可视化与建模方法

肖光恩

武汉大学经济与管理学院世界经济系

武汉大学经济发展研究中心

- 主要内容
  - 新冠疫情全球大流行的基本特征
  - 新冠疫情数据来源
  - 新冠疫情数据可视化的方法
  - 新冠疫情数据的建模方法

一、新冠疫情全球大流行的基本特征

- 新冠疫情传播的全球性
  - 2019年12月31日：确认未明肺炎病为新冠病毒
  - 2020年1月13日:泰国报告一例新冠病例
  - 1月22日：确认存在人际传播
  - 1月23日：武汉市封城
  - 1月30日: 新冠全球疫情等级非常高
  - 2月29日：全球关注的公共卫生事件（PHEIC）
  - 3月11日：全球大流行（Pandemic）
  - 3月13日：美国宣布进入“紧急状态”
  - 8月14日：确诊2073万人；确诊死亡：大于75万人；确诊康复：大于1278万人，196个国家地区

- 新冠疫情流行的长期性
  - 科学家观点：潜伏期长，传播方式多样；控制难，传播快
  - 经济学观点：医疗挤兑，医疗物质短缺，全球医疗物质流通困难
  - 政治学家观点：保护主义，民粹主义，治理赤字

- 新冠疫情流行的破坏性
  - 生产要素视角：要素缺失（劳动力死亡），要素流动（居家隔离），要素配置（全球市场隔离）
  - 市场视角：供给中断，需求中断，供给生态破坏
  - 产业视角：产业脱钩——生产链、价值链、供应链

一、新冠疫情全球大流行的基本特征

- 新冠疫情防控政策的阻隔性
  - 封闭管理（lockdown）：
  - 旅行与贸易限制（travel ban）：全球217个目的地
    - 关闭边境（97个）：45%
    - 旅行限制（39个）：18%
    - 终止航班（65个）：30%
    - 特定限制（自我隔离与签证限制，16个）：7%

- 新冠疫情全球治理的赤字性：去全球化（全球主义）vs.本土化（亲民粹主义）
  - UN:“缺位”
  - WTO：“缺席”
  - G20：“缺能”
  - 中美两国：“缺和”

二、新冠疫情数据来源

- 国内渠道：
  - 新华网：http://my-h5news.app.xinhuanet.com/h5activity/yiqingchaxun/index.html
  - 凤凰网：https://news.ifeng.com/c/special/7uLj4F83Cqm
  - 新浪网：https://news.sina.cn/zt_d/yiqing0121
  - 国家卫健委：
  - 专家：清华大学教授柯惠新教授公众号：“六人团队”

- 国外渠道：
  - WHO：https://covid19.who.int/
  - Tableau: https://www.tableau.com/covid-19-coronavirus-data-resources
  - Johns Hopkins:https://coronavirus.jhu.edu/map.html
  - Worldometers: https://www.worldometers.info/coronavirus/
  - CDC: https://www.cdc.gov/covid-data-tracker/#cases

三、新冠疫情数据可视化的方法

- 新冠疫情数据可视化的目的之一：单变量
  - 新冠疫情传播的趋势特征：
    - 总体趋势
    - 个体趋势
  - 新冠疫情传播的差异特征
    - 横向差异
  - 新冠疫情传播追踪特征
    - 增长模型
    - 面板模型
  - 新冠疫情传播的溢出特征
    - 截面溢出特征
    - 面板溢出特征
    - 动态溢出特征

三、新冠疫情数据可视化的方法

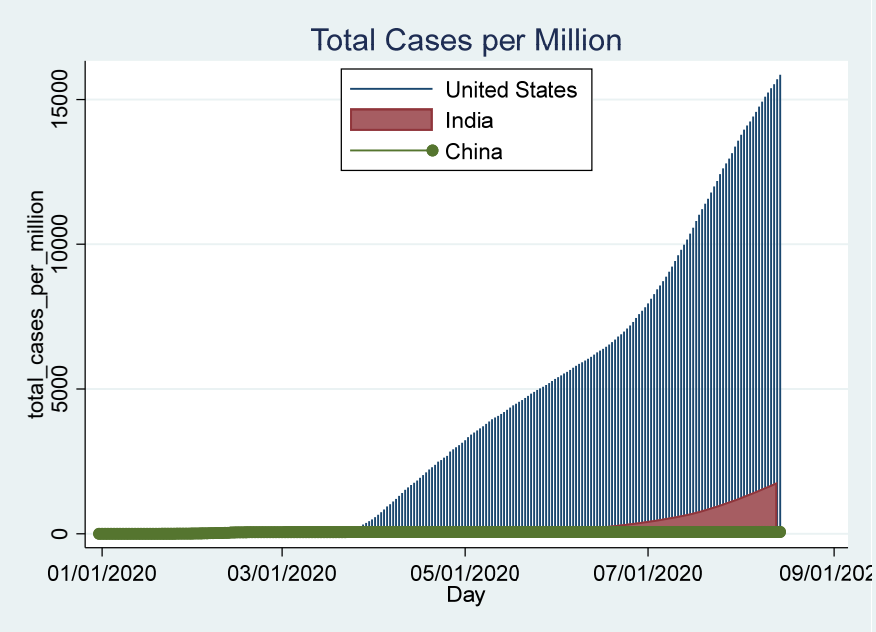
- 新冠疫情数据可视化的目的之二：双变量
  - 截面数据：
    - 共变关系
    - 因果关系
  - 面板数据
    - 共变关系
    - 因果关系
  - 空间数据
    - 被解释变量的空间溢出
    - 解释变量的空间溢出
    - 残差项的空间溢出
    - 动态空间溢出
    - 复合空间溢出
  - 流行病学数据
    - 生存数据

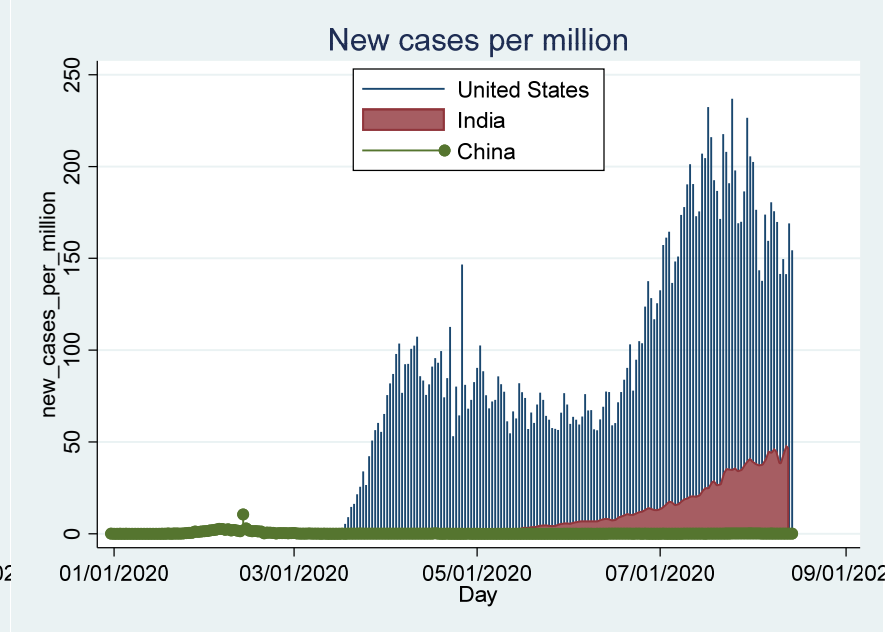
三、新冠疫情数据可视化的方法

- 新冠疫情绝对水平数据
  - 确诊总病例：total_cases
  - 新增病例：new_cases
  - 确诊总死亡病例：total_deaths
  - 确诊新增死亡病例：new_deaths
- 新冠疫情相对水平数据
  - 确诊总病例比率：total_cases_per_million
  - 确诊新增病例比率： new_cases_per_million
  - 确诊总死亡病例比率：total_deaths_per_million
  - 确诊新增死亡病例比率：new_deaths_per_million
- 新冠疫情救治数据
  - 新增检测人数：new_tests
  - 总检测人数：total_tests
  - 新增检测人数比率： total_tests_per_thousand
  - 总检测人数比率： new_tests_per_thousand
  - 检测阳性比率： positive_rate
- 新冠疫情成因数据（国家与个体层面）
  - 政府严格指数：stringency_index
  - 人均GDP（极端贫困指数、人口密度）：gdp_per_capita（extreme_poverty，population_density)
  - 医院床位数（洗手设施）：hospital_beds_per_thousand（handwashing_facilities)
  - 年龄组（吸烟、糖尿病）： aged_65_older（smokers，diabetes)

三、新冠疫情数据可视化的方法

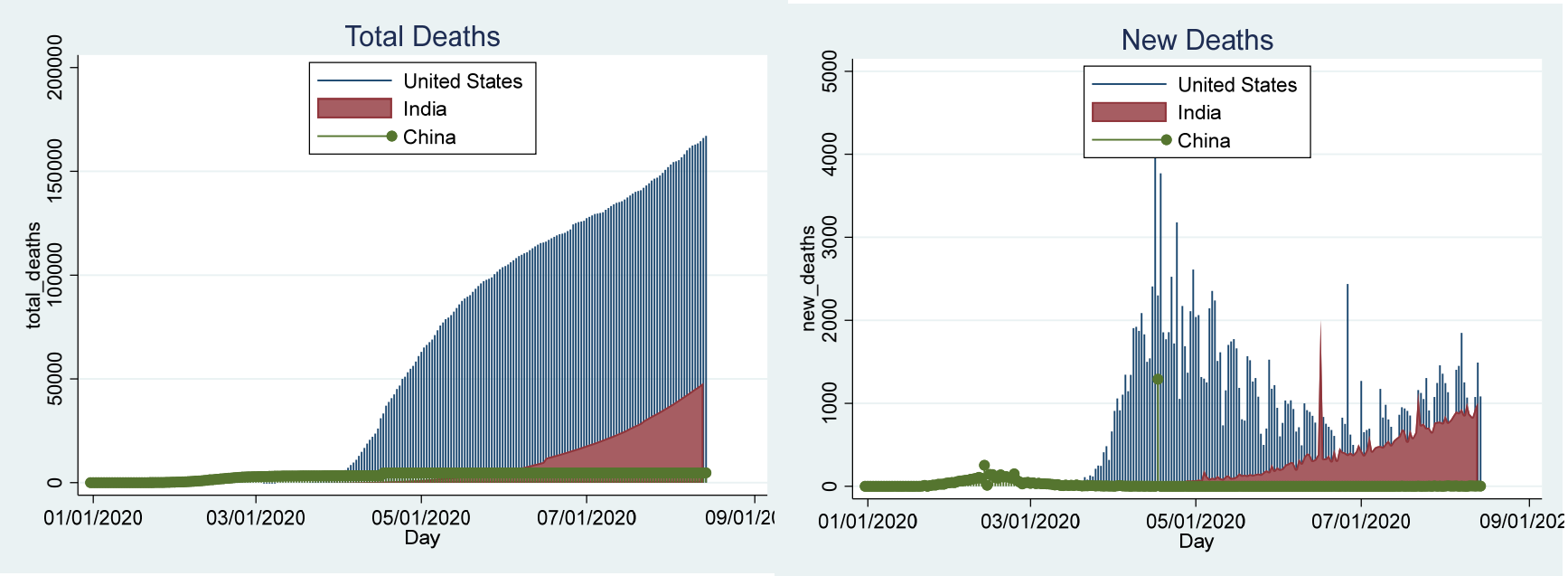
◘ 新冠疫情数据可视化的示例





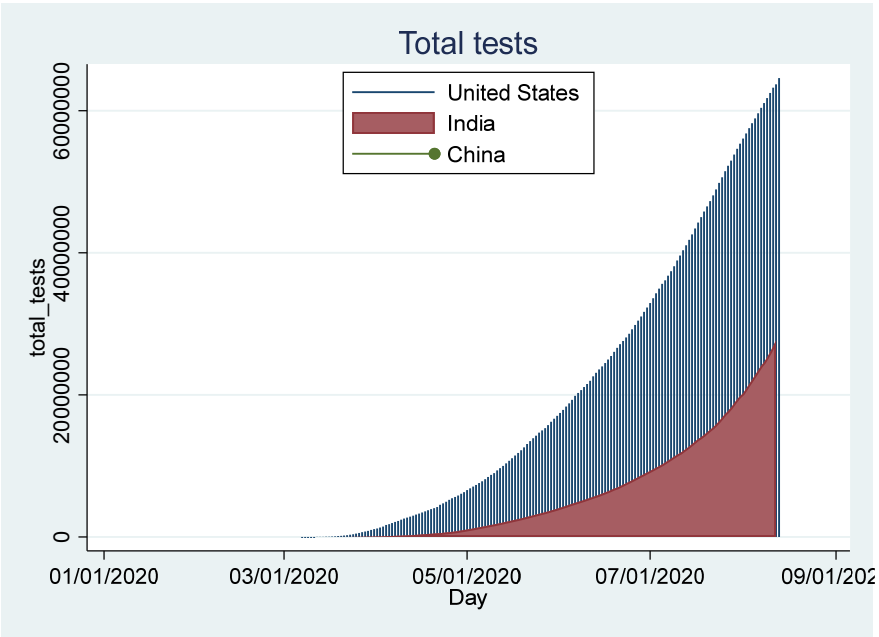
三、新冠疫情数据可视化的方法

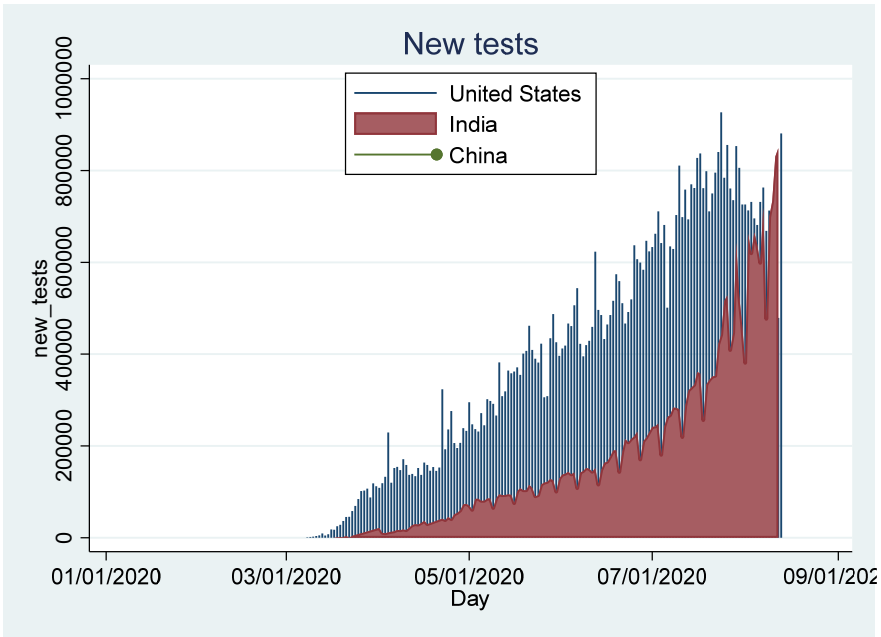
◘ 新冠疫情数据可视化的示例



三、新冠疫情数据可视化的方法

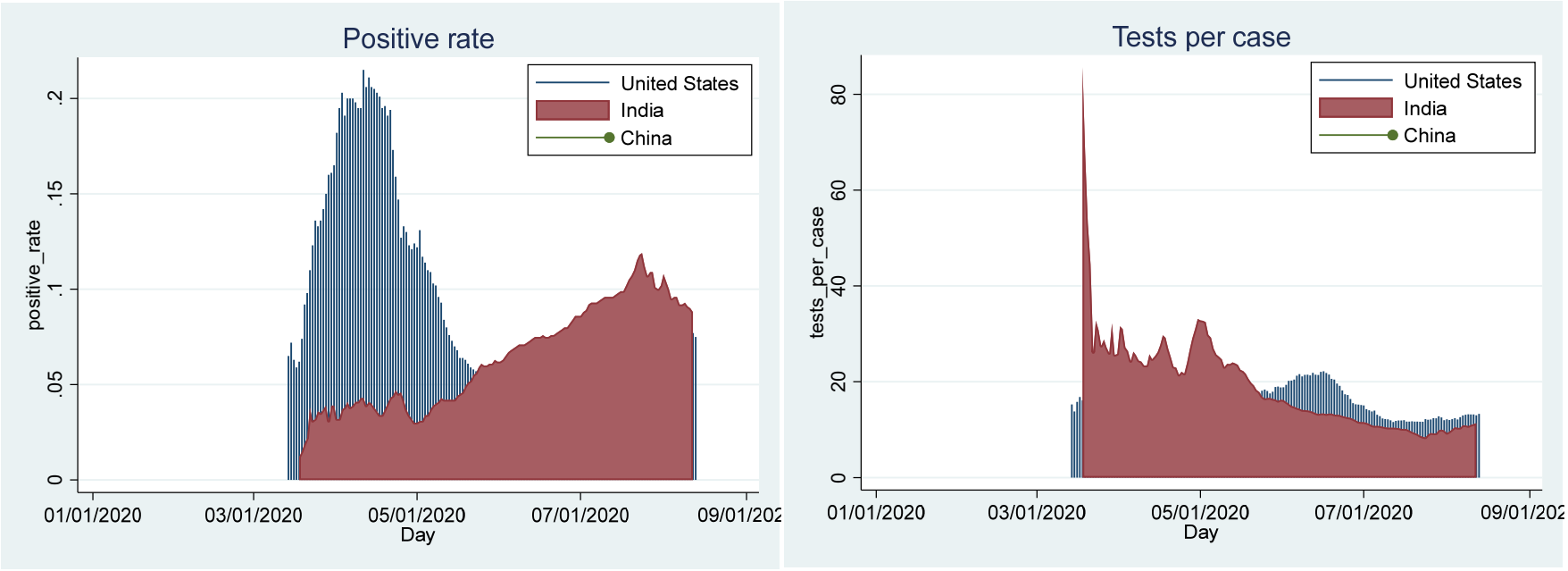
◘ 新冠疫情数据可视化的示例





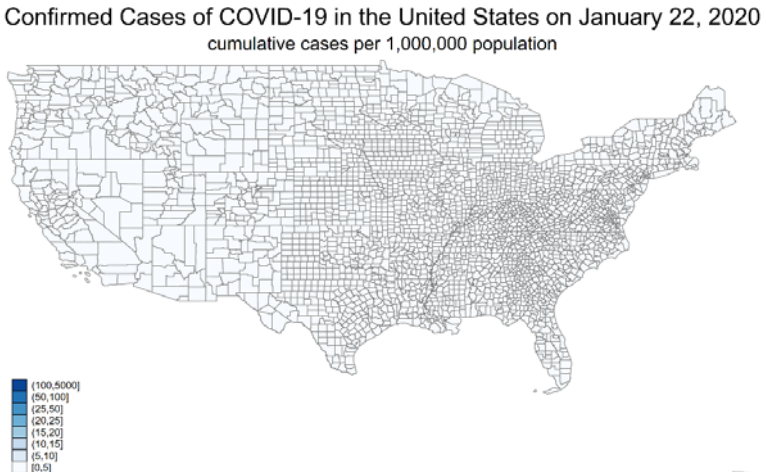
三、新冠疫情数据可视化的方法

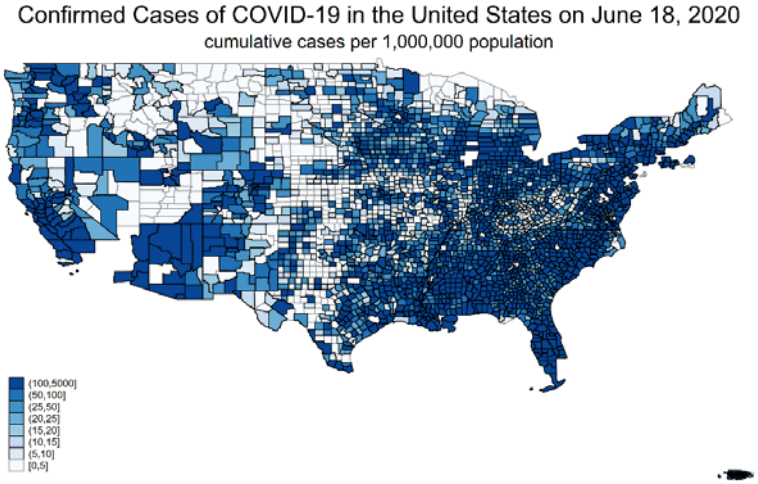
◘ 新冠疫情数据可视化的示例

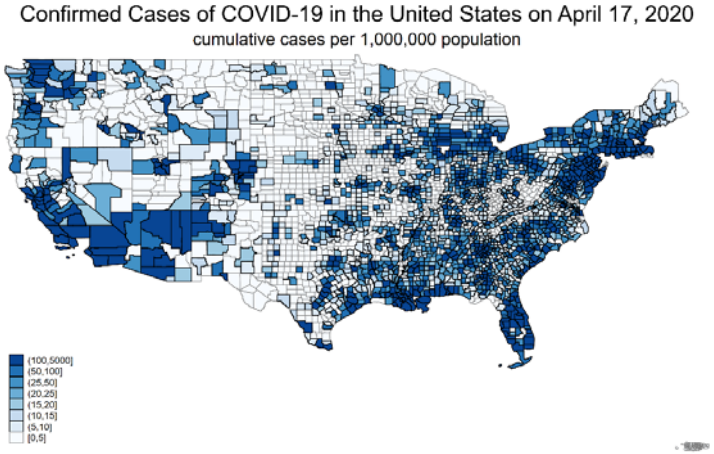


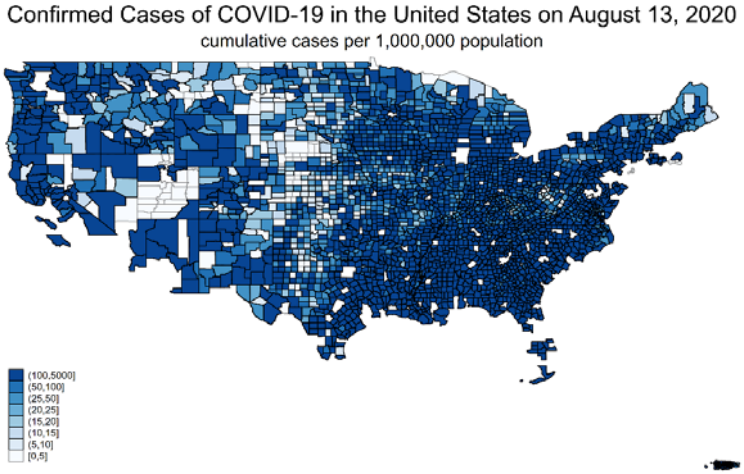
三、新冠疫情数据可视化的方法

- 新冠疫情数据可视化的示例







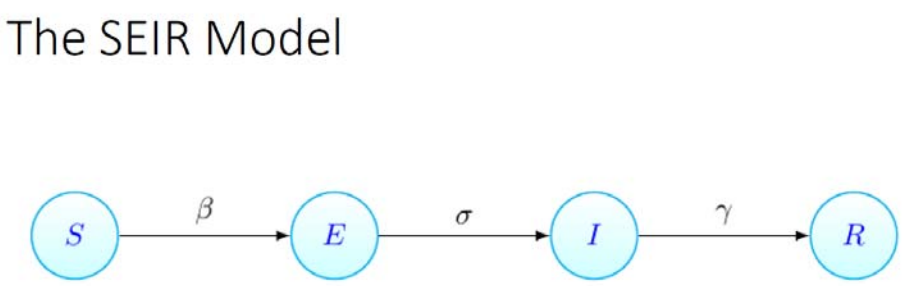


四、新冠疫情数据的建模方法

- 新冠疫情数据建模的思路：
  - 关注被解释变量自身特征：
    - 确诊病例研究感染者的生存状态：
    - 研究易感染人群、感染人群、移除（治愈）人群三者之间的转换关系
  - 关注被解释变量或解释变量的溢出效应
    - 空间效应模型
    - 动态空间模型
    - 时间-空间模型
    - 来源-目的空间模型
  - 关注被解释变量生态系统
    - 系统方程
    - 状态转换模型
    - VAR/面板VAR模型
  - 关注被解释变量的增长趋势
    - 时间序列模型
    - 增长模型、
  - 政策效应的计量分析
    - DID
    - Propensity Score

四、新冠疫情数据的建模方法

- 新冠疫情数据的建模方法：SEIR模型
  - 基本设定：S(susceptible)-E(Exposed)-I(Infected)-R(Recovered)
  - S:易感染人数
  - E:在病毒流行期间面临病毒威胁但仍没有感染的人数
  - I:已经感染的人数
  - R:已经治愈的人数
  - N:总人数（总体）
- SEIR模型的假设
  - 在疫情流行期间的任何时间点，总体中的个体只能分属于三个群体中的一种
  - 个体一但治愈便不再被感染
  - 总体中的自然出生率与自然死亡率是平衡的
  - 总体在短期内不会发生变化，且不存在移民或居民迁移
- 变量设定
  - s: S/N
  - e: E/N
  - i: I/N
  - r: R/N
- 四个变量之间的逻辑关系
  - S----(β)---E---(σ)---I---(γ)---R
  - N=S+E+I+R
  - s+e+i+r=1



四、新冠疫情数据的建模方法

- 参数设定及其含义
  - β：每天每个感染者通过接触感染的人数，如 $\beta = 1/2$（即感染者每2天感染1个人）
    - 通常用一个感染者通过接触感染所需要平均期限（天数）的倒数来表示。
    - 平均天数通常是个外生变量，由研究者根据疾病知识人为设定
  - $\gamma = R/\mathrm{I}$：感染人群被治愈的比重。
    - 通常用感染平均时间的倒数来表示，
    - 如，$\gamma = 1/3$ ， 即平均每3天有一个人被治愈
    - 它是一个外生变量，由研究者根据疾病规律来设定
  - σ =传染率，即一个感染者其病毒传染的概率
    - 通常用感染者在感染时的平均周期的倒数表示，它与传染率的关系为：$\frac{1}{\sigma}$ = 感染时的平均周期。
    - $\sigma = 0.25 = 1/4$，即感染者在感染状态的平均时间为4天，或者解释为25%的被感染者或面临感染的人在感染状态需要的平均时间（用天表示）
    - 它是一个外生变量，研究者自行设定

四、新冠疫情数据的建模方法

- SEIR的系统方程及其求解
  - 系统微分方程（方程都是非线性关系）
  - 四个变量：s, e, i, r
  - 三个参数：β， σ， γ
  - 系统方向有四个方程：

$$\frac{dS}{dt} = -\frac{\beta SI}{N}$$
$$\frac{dE}{dt} = \frac{\beta SI}{N} - \sigma E$$
$$\frac{dI}{dt} = \sigma E - \gamma I$$
$$\frac{dR}{dt} = \gamma I$$

  - 系统方程的求解：数值法与非数值法，常用求微分的方法

四、新冠疫情数据的建模方法

- SEIR模型预测的决定因素：假设总体为1
  - 当感染群体比例（i）上升时，治愈群体比例（r）会上升，易感染人群比例（s）会下降
  - 不同群体比例变化波动取决于三个参数β， σ， γ的设定
  - **关于β**
    - 当一个感染者被感染所需要的平均时间增加，即β值变小时，则每天感染率下降且会变量更平坦。这说明保持**社交距离**就很重要，它很降低被感染率，增加了被感染所需要的平均时间长度。
  - **关于γ**
    - 当感染平均期限增加，即γ值变小时，则每天的感染率会上升，这说明感染的**治愈方法或方案**就非常重要，即治愈效果的提高有利于减少感染所需要的平均时间。
  - **关于σ（感染率）**
    - 当处在感染状态时的平均期限增加，即σ的值变小时，则每天感染率开始上升很慢，但随后会增长得更快。这说明保持**社交距离**很重要，它能确保处在感染状态的人不能传播病毒，可以为政府部门换取更多的时间来提高公共防卫措施。
- Stata的估计命令： EPIMODELS： epi_seir； epi_sir

四、新冠疫情数据的建模方法

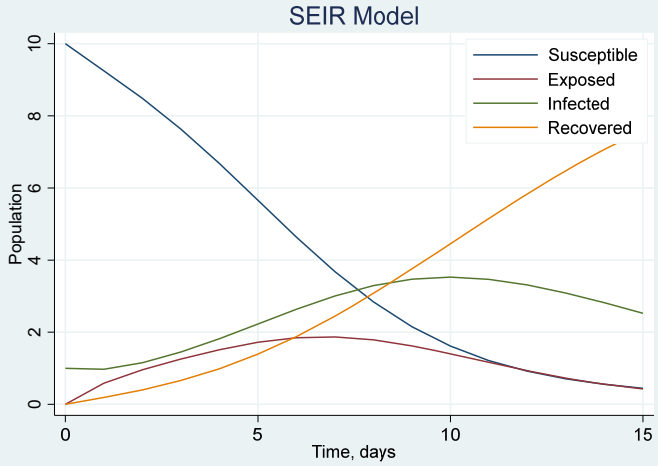
- 新冠疫情数据的建模方法：SEIR模拟结果取决地参数
  - 当15天之内有10个易感染人数和1个感染者

```
. epi_seir , days(15) beta(0.9) gamma(0.2) sigma(0.5) susceptible(10) infected(1)
```

SEIR Model

Population	t0	t10	t15
Susceptible	10	2	0
Exposed	0	1	0
Infected	1	4	3
Recovered	0	4	8
Total	11	11	11

The maximum size of the infected group 3.527 is reached on day 10 of the simulation.

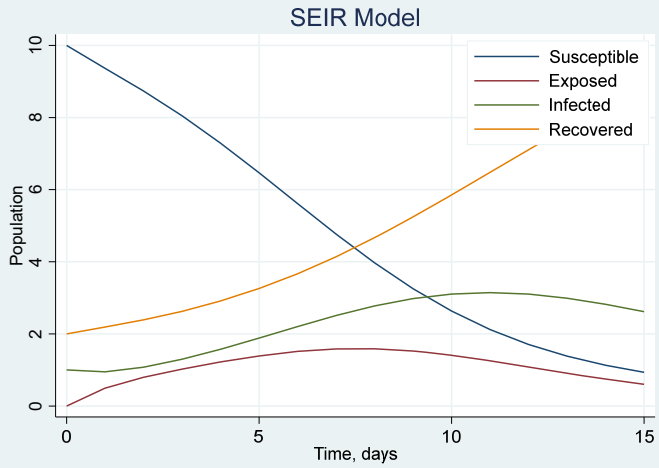



```
. epi_seir , days(15) beta(0.9) gamma(0.2) sigma(0.5) susceptible(10) infected(1) recovered(2) clear
```

SEIR Model

Population	t0	t11	t15
Susceptible	10	2	1
Exposed	0	1	1
Infected	1	3	3
Recovered	2	6	9
Total	13	13	13

The maximum size of the infected group 3.145 is reached on day 11 of the simulation.



四、新冠疫情数据的建模方法

- 新冠疫情数据的建模方法：SEIR模拟结果取决地参数
  - 当15天之内有10个易感染人数和1个感染者

```
. epi_seir , days(15) beta(0.9) gamma(0.2) sigma(0.5) susceptible(10) infected(1) recovered(2) clear
```

SEIR Model

Population	t0	t11	t15
Susceptible	10	2	1
Exposed	0	1	1
Infected	1	3	3
Recovered	2	6	9
Total	13	13	13

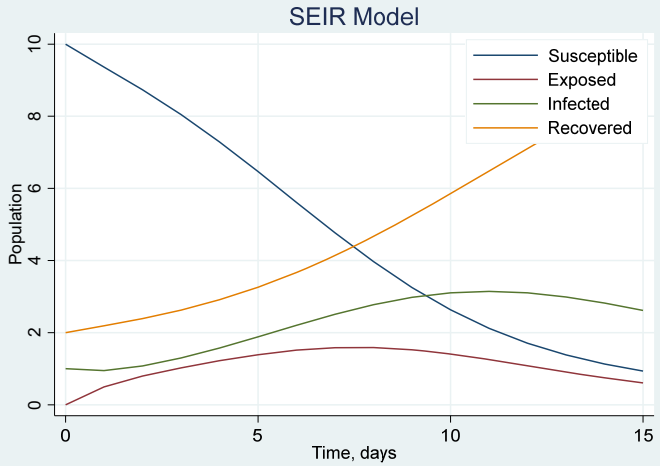
The maximum size of the infected group 3.145 is reached on day 11 of the simulation.

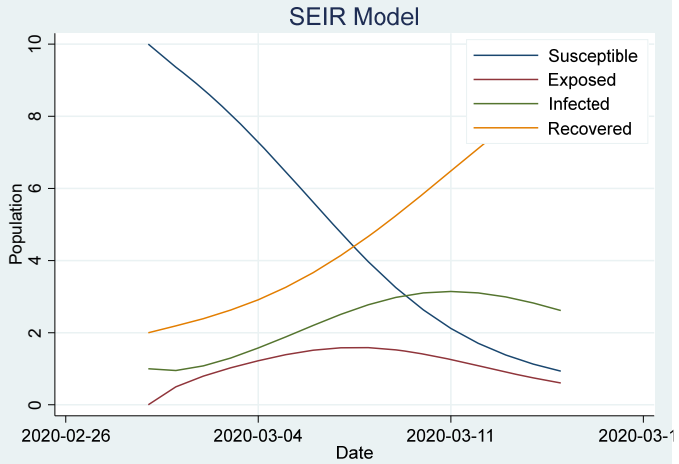
```
. epi_seir , days(15) day0("2020-02-29") beta(0.9) gamma(0.2) sigma(0.5) susceptible(10) ///
> infected(1) recovered(2) clear
```

SEIR Model

Population	2020-02-29	2020-03-11	2020-03-15
Susceptible	10	2	1
Exposed	0	1	1
Infected	1	3	3
Recovered	2	6	9
Total	13	13	13

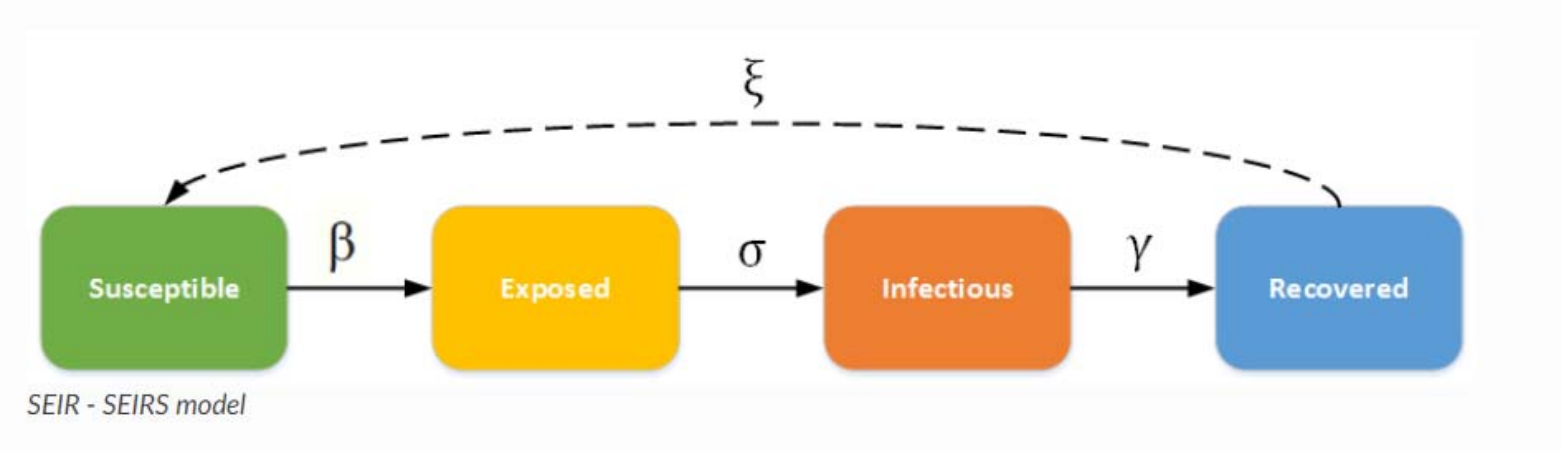
The maximum size of the infected group 3.145 is reached on day 11 (2020-03-11) of the simulation.





四、新冠疫情数据的建模方法

- 新冠疫情数据的建模方法：SEIR模型的变形
  - SER:
  - SEIRS:



  - SEIR without vital dynamics
  - SEIR with vital dynamics

- https://www.idmod.org/docs/emod/hiv/model-seir.html

Thank You!