Max Binder

PREDICTING THE PROBABILITY OF INFECTION FOR AN AGENT-BASED MODEL FROM VIRUS ATTRIBUTES

AP Statistics Final Project

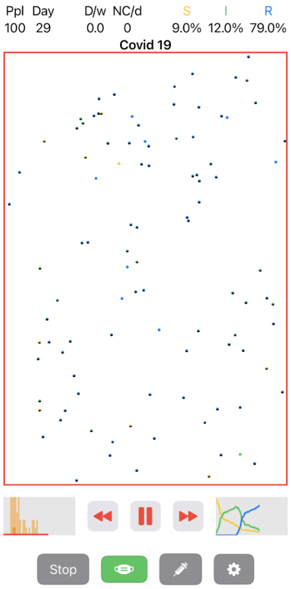
Stanford Online High School ★ May 2025

Introduction

After COVID-19, many people argued against the effectiveness of wearing masks, getting vaccinated, etc. To answer this question for myself, I created an application that I call Virusology, which simulated the spread of a virus in a closed population of little spherical people. However, I learned fast that translating our complicated biology into simple code was not so easy. In my application, when two spheres come close to each other, one of which is infected, what is the probability that the second one gets infected? In real life, viruses have their own attributes, and my challenge was translating those attributes into the probability of infection. This paper will describe a method to estimate the probability of infection (for the purpose of the Virusology application) based on real-life attributes of a virus and a population density.

Virusology Application Description

Virusology is a simulation platform developed to visualize and study the spread of viruses in a simplified digital environment. It uses an agent-based model where each agent represents an individual that moves randomly and interacts with others according to a set of rules. Agents may become infected, recover, or remain susceptible depending on simulation settings. The user can control key parameters such as the virus's contagiousness, recovery time, fatality rate, and the overall population density. The simulation makes it easy to see how infections grow or die out under different conditions, helping users better understand how various factors influence the course of an outbreak.



Ppl	Day	D/w	NC/d	S	I	R
100	29	0.0	0	9.0%	12.0%	79.0%

Covid 19

Figure 1: Screenshot of Virusology Application.

The Challenge

While designing the simulation, I encountered a major difficulty: how to relate the simulation's mechanics to real-world virus statistics. In particular, I needed a way to use R_0 values to drive the infection dynamics of the simulation. Since infections in Virusology occur through random contact between agents, I had to determine what value of “probability of infection” would produce an R_0 close to what we see in reality. But this probability is affected by many things, including the specific virus attributes, how often agents move and collide, and how crowded the simulation space is. This made it difficult to guess or calculate the right value without testing. The challenge became finding a reliable, data-driven way to translate a known R_0 into an appropriate infection probability for the simulation.

Methodology

To solve this, I will use ordinary least squares (OLS) regression. OLS regression is a technique used to estimate the relationship between a dependent variable and one or more independent variables. It identifies the linear equation that minimizes the sum of the squared differences between the observed values and the values predicted by the model. OLS assumes a linear relationship, constant variance of errors, independence of observations, and normally distributed errors.

In this case, the dependent variable is:

- $y = \mathrm{P}_{\text{infection}}$: This is the probability that during an interaction between an infected agent and a susceptible agent, the virus is transmitted.

The independent variables are:

- $x_1 =$ **Mortality Rate**: This is the probability that an infected individual will die;
- $x_2 =$ **Incubation Period**: This is the time between the exposure to the virus and onset of symptoms;
- $x_3 =$ **Presymptomatic Period**: This is the subset of the incubation period during which the infected person has no symptoms but is already contagious;
- $x_4 =$ **Contagious Period**;
- $x_5 =$ **Population**: Since the simulation always takes place in spaces of equal size, population is effectively the *population density*;
- $x_6 = R_0$: This is the average number of secondary infections caused by one infected individual in a fully susceptible population

The incubation period, presymptomatic period, and contagious period collectively make up the **Recovery Time**.

The linear relationship between the dependent and independent variables is described by the following system of linear equations:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_6 x_{6,i} + \varepsilon_i \qquad (i = 1, \ldots, N),$$

where y_i is the realization of the dependent variable, $x_{j,i}$ is a realization of the independent variable x_j ($j = 1, \ldots, 6$), and N is the number of test runs.

Having the above system of equations, OLS aims to solve the following optimization problem:

$$\min_{\beta_0, \ldots, \beta_6} \sum_{i=1}^{N} \left[\left(y_i - \left(\beta_0 + \sum_{j=1}^{6} \beta_j x_{j,i} \right) \right)^2 \right]$$

I will conduct the regression analysis using the Python programming language[1].

[1] See Bibliography, Source 2 for Python.

Data

To use Ordinary Least Squares Regression, I need data. I designed a framework in which the simulation runs many times using different combinations of virus traits and environmental settings. In each simulation, I fixed certain variables—such as death rate, recovery time, and density—and varied the probability of infection. I then calculated the average number of people infected by each agent, which serves as the simulated R_0. By running multiple iterations for each setting, I was able to reduce the impact of random variation and obtain reliable averages. This gave me a dataset containing the input values and their resulting R_0.

Specifically, I varied all values and calculated R_0:

- Mortality Rate: 0.0006, 0.001, 0.025
- Incubation Period: 2, 5
- Presymptomatic Period: 1, 2
- Contagious Period: 5, 7, 10
- Population: 50, 100, 250
- $\mathrm{P}_{\text{infection}}$: 0.01, 0.025, 0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5

Some data is summarized in the table below[2].

x_1	x_2	x_3	x_4	x_5	$\mathrm{P}_{\text{infection}}$	R_0
0.01	2	1	7	250	0.05	$2.3\ldots$
0.025	5	2	5	100	0.01	$1.2\ldots$
0.0006	2	1	10	100	0.01	$1.8\ldots$
0.001	5	2	7	250	0.01	$3.6\ldots$
0.025	2	1	5	50	0.05	$1.7\ldots$
0.025	2	1	10	50	0.05	$3.0\ldots$
0.001	5	2	10	250	0.075	$3.7\ldots$
0.025	5	2	10	100	0.025	$3.8\ldots$

Table 1: Sample data (8 randomly-selected rows out of 972).

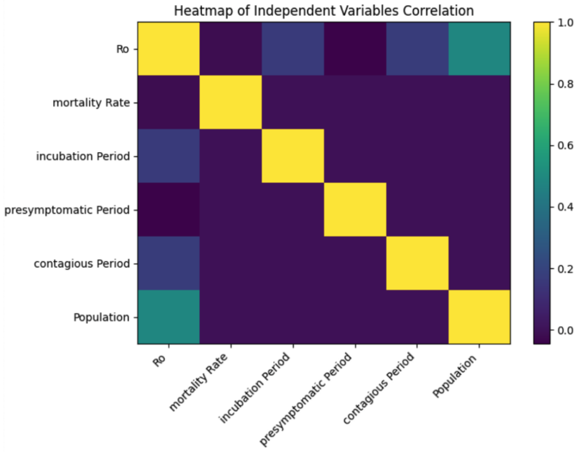


Figure 2: Pairwise correlation between indepedent variables.

[2] See Bibliography, Source 1 for complete data set.

For example, from the data above, we can see that the correlation between the population and R_0 is ≈ 0.6.

Results

The results are summarized below. The ordinary least-squares (OLS) regression analyzed the relationship between the probability of infection and six indepedent variables: mortality rate, incubation period, presymptomatic period, contagious period, population, and R_0 (x_1 through x_6). The model explains approximately 45.7% of the variance in infection probability ($R^2 = 0.457$), and the overall regression is statistically significant ($F(6, 965) = 135.5, p < 0.001$).

Dep. Variable:	$\mathrm{P}_{\mathrm{infection}}$	R^2:	0.457
Model:	OLS	Adj. R^2:	0.454
Method:	Least Squares	F-statistic:	135.5
No. Observations:	972	Prob (F-statistic):	2.33×10^{-124}
DF Residuals:	965	Log-Likelihood:	654.14
DF Model:	6	AIC:	$-1294.$
Covariance Type:	nonrobust	BIC:	$-1260.$

	coef	std err	t	$P > \|t\|$	$[0.025$	$0.975]$
const	0.1844	0.004	46.413	0.000	0.177	0.192
R_0	0.1350	0.005	28.508	0.000	0.126	0.144
Mort rate	0.0022	0.004	0.559	0.576	-0.006	0.010
Inc Per	-0.0222	0.004	-5.487	0.000	-0.030	-0.014
Presymp P	0.0060	0.004	1.519	0.129	-0.002	0.014
Cont Per	-0.0232	0.004	-5.718	0.000	-0.031	-0.015
Popul	-0.0657	0.005	-14.303	0.000	-0.075	-0.057

Omnibus:	94.703	Durbin-Watson:	0.691
Prob(Omnibus):	0.000	Jaque-Bera (JB):	120.736
Skew:	0.837	Prob(JB):	6.06×10^{-27}
Kurtosis:	3.420	Cond. No.	1.84

Table 2: OLS regression results.

Among the predictors, incubation period; contagious period; population; and R_0 were statistically significant at the 0.05 level.

Notably:

- R_0 had a positive and highly significant effect ($\beta = 0.1350, p < 0.001$), suggesting that higher transmission rates increase the probability of infection;
- Incubation period ($\beta = -0.0222$, $p < 0.001$) and contagious period ($\beta = -0.0232$, $p < 0.001$) both showed significant negative associations, implying that longer durations are associated with lower infection probability;
- Population also had a strong negative effect ($\beta = -0.0657$, $p < 0.001$), potentially reflecting dilution or saturation effects in larger populations.

The mortality rate and presymptomatic period did not reach statistical significance ($p = 0.576$ and $p = 0.129$, respectively).

Conclusion

In this project, I created a dataset by running nearly 1000 simulations in my Virusology application, each with different virus traits and population settings. I then used ordinary least squares regression to model how these traits—such as R_0, incubation period, and population size—affect the probability of infection between agents.

The regression results showed that R_0, incubation period, contagious period, and population size were all statistically significant predictors of infection probability, with the model explaining about 46% of the variation. This provides a practical way to estimate the infection probability in a simulation based on real virus attributes:

$$\begin{aligned}\mathrm{P}_{\text{infection}} = 0.1844 + 0.1350R_0 + 0.0022[\text{Mortality Rate}] - 0.0222[\text{Incubation Period}] + \\ + 0.0060[\text{Presymptomatic Period}] - 0.0232[\text{Contagious Period}] - 0.0657[\text{Population}]\end{aligned}$$

While the OLS model was effective, it had limitations. The relationship between variables may not be purely linear, and using nonlinear regression or more complex models could improve accuracy. Nonetheless, this project demonstrates how simulation and statistical techniques can work together to better understand and model the spread of a virus.

Bibliography

1. Binder, Max. (2025). *Virusology Simulation Dataset* [Google Sheets].
   https://docs.google.com/spreadsheets/d/1jxzIcr7R1W9fLrcCADtaq0vRN8-1New7dHe75nzuWoM/
2. Binder, Max. (2025). *OLS Regression Analysis for Virusology Project* [Python code]. Google Colab.
   https://colab.research.google.com/drive/17fJfQvm458tc2E1NLc211VqVD9JYhaqW
3. Binder, Max. (2025). *Virusology* [ABM of the spread of a Virus]. Self-published.
4. Kerr C C, Stuart R M, Mistry D, et al. *Covasim: An agent-based model of COVID-19 dynamics and interventions.* PLOS Computational Biology. 2021;17(7):e1009149. doi:10.1371/journal.pcbi.1009149
5. Steinhöfel K K, Heslop D, MacIntyre C R. *Agent-Based Models of Virus Infection.* Current Clinical Microbiology Reports. 2025;12(1):2. doi:10.1007/s40588-024-00238-5