Exam3

Zuojun Gong

Sunday, May 03, 2015

Introduction

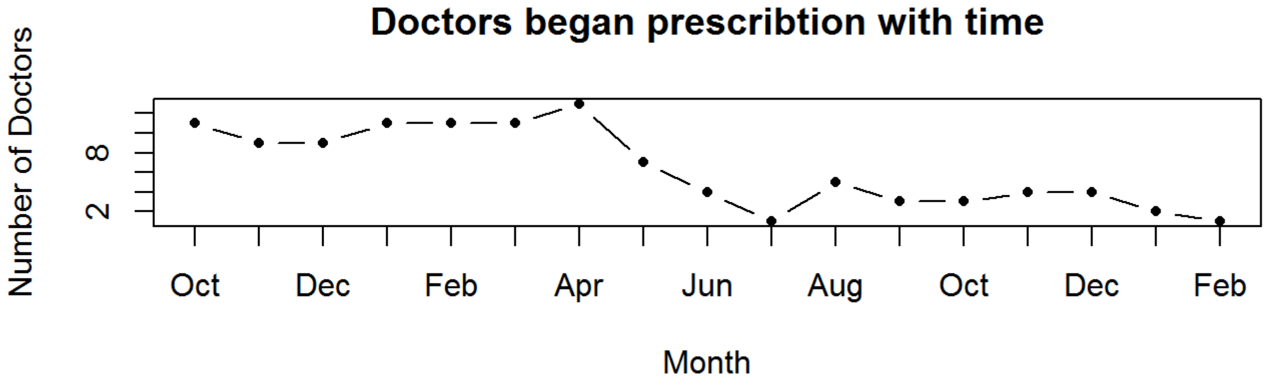
Ideas are a powerful source that can push humanity forward. Great ideas starts trivial, but will eventually spread like virus. In modern society, we want to capture how exactly ideas spread. In particular, we want to examine how the new antibiotic tetracycline spreads among doctors in four local communities in the state of Illinois back in 1950s. We want to focus on how the doctors start to adopt the new idea of using tetracycline, and the relationship between their innovation and their medical school education, medical journal read, medical conferences attended, and how the other doctors within the network react to the new antibiotics. Some argue that ideas and innovation spread from person to person through direct contact, and we want to test this hypothesis through statistical models and casual inference.

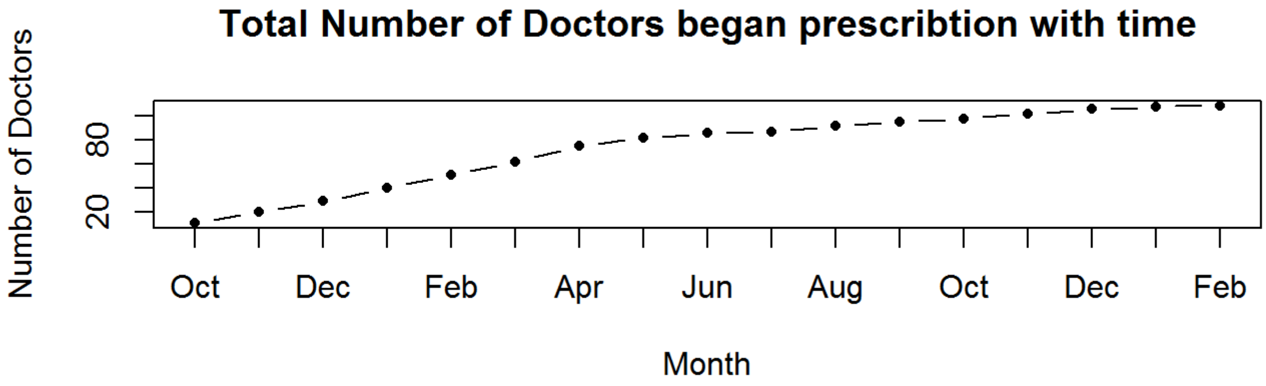
Modeling

Discovering underlying patterns

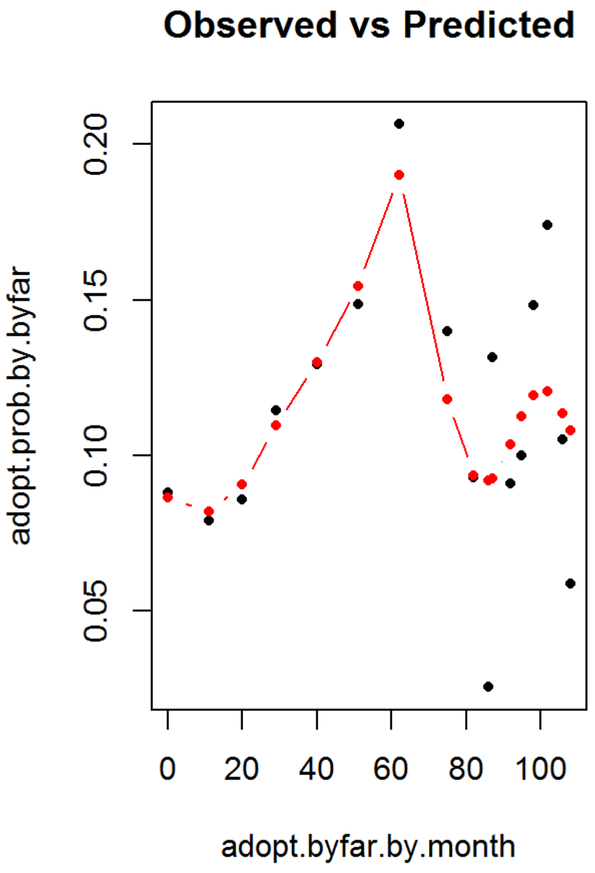
We have information from 264 doctors from four different communities across 17 months starting from November 1953. We know about their date of adopting the new antibiotics, years when they attended medical school, whether they attend medical conference lately, the number of medical journals they read within the time period. We also know who they usually spend their free time with, if the discuss medicine socially, and other variables. We also have a record of connections between each doctors within the study in a binary matrix. We want to use these information to help us to infer how the innovation ideas spreads.

In order to observe how the new idea spread amongst the doctors in these four communities, we want to focus on the doctors that we have information about their adoption date of the drug. Then we look at the graph of the number of doctors adopting the new drug over time, and the total number of doctors adopting the drug over time.

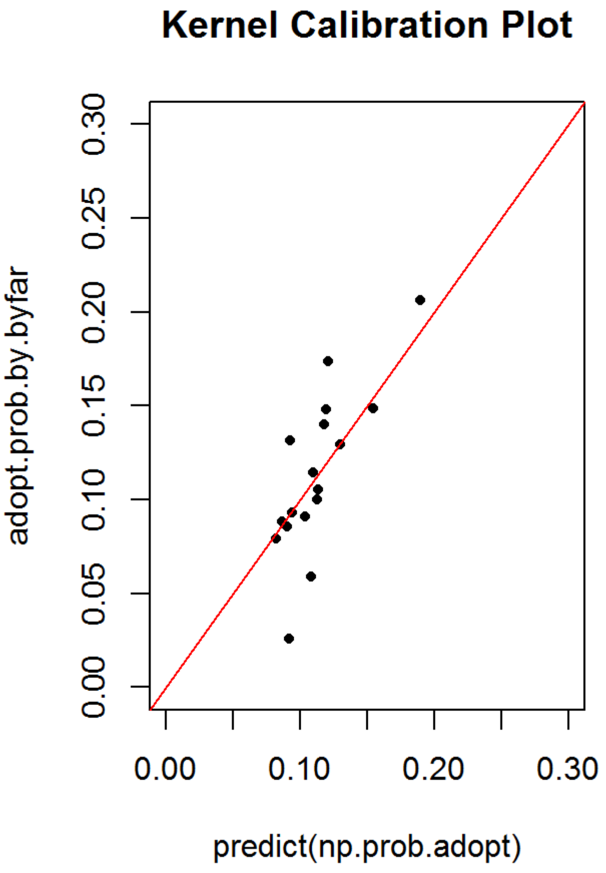




We can observe from the graphs above that starting from the very first month all the way to half a year later, the number of doctors prescribing the drug is increasing. However after July 1954 the number of doctors who begun to adopt the drug dropped to a relatively steady and low level (this is also reflected on the total number of doctors prescribing). We can see that the doctor's in the local community is experiencing a “saturated” effect: number of doctors prescribing decreases after certain period of time as the total number of doctors prescribing increases. Therefore we want to explore the potential relationship between the probability of a doctor adopting the new drug and the total number of doctors already prescribing the drug.

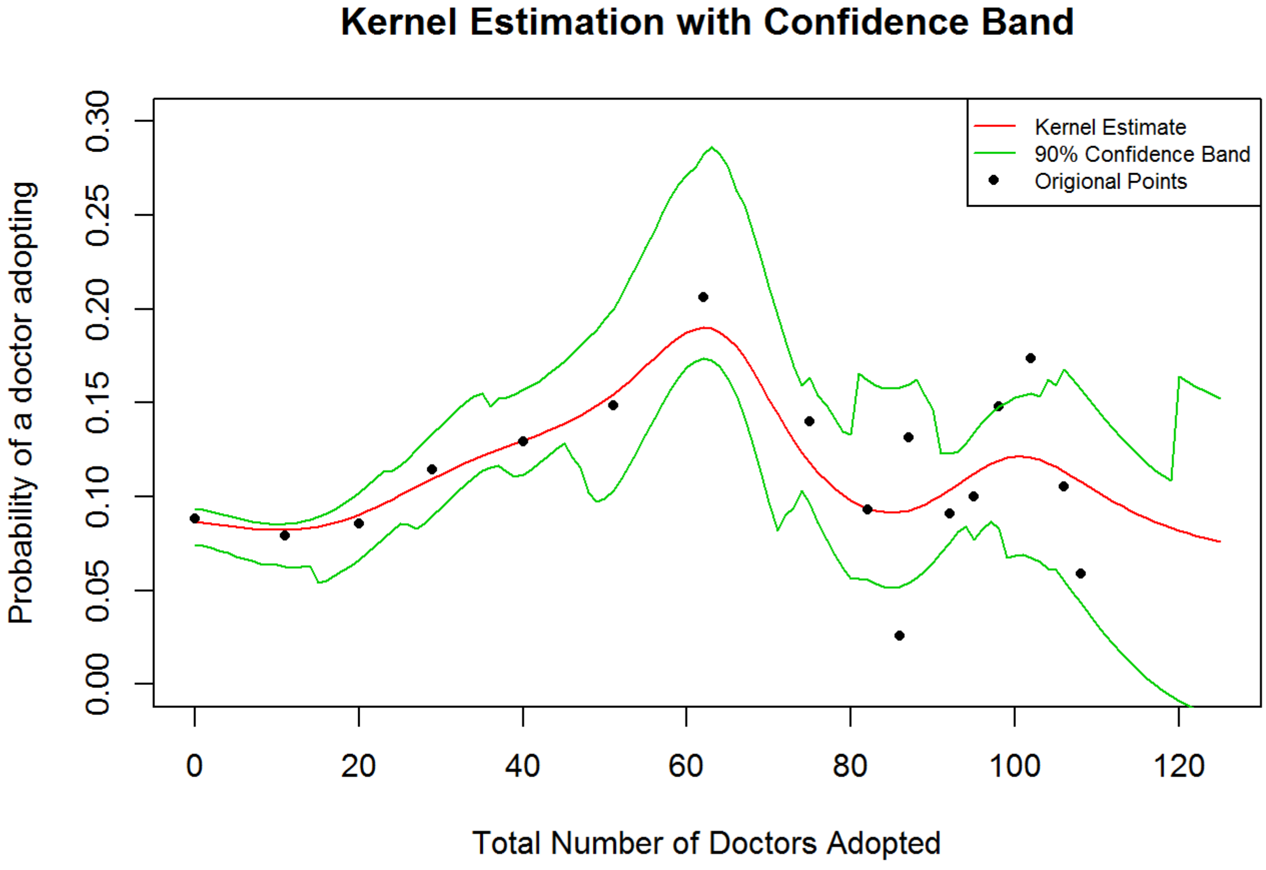


Observed vs Predicted



Kernel Calibration Plot

Under our assumptions, at each given month t, the probability of a doctor adopting the drug is computed by the number of doctors adopting the drug at month t over the total number of doctors that have not adopted the new drug this month. We want to model these probabilities against the total number of doctors already prescribing the new antibiotics at this month. We proposed two models, kernel regression and generalized linear models with logistics regression. We compare both models by their MSEs. Kernel model's cross-validated MSE is 0.00169 and that for GLM is 0.002. Besides the better goodness of fit, we conclude that Kernel regression is more appropriate in this situation because we do not need to assume any distribution or ordering of our observed value, which appears to be in a curve. We then plot the model's estimated probabilities against the observed values, and it appears to be a good fit. With the low MSE and appropriate goodness of fit, we decided to proceed with kernel regression.

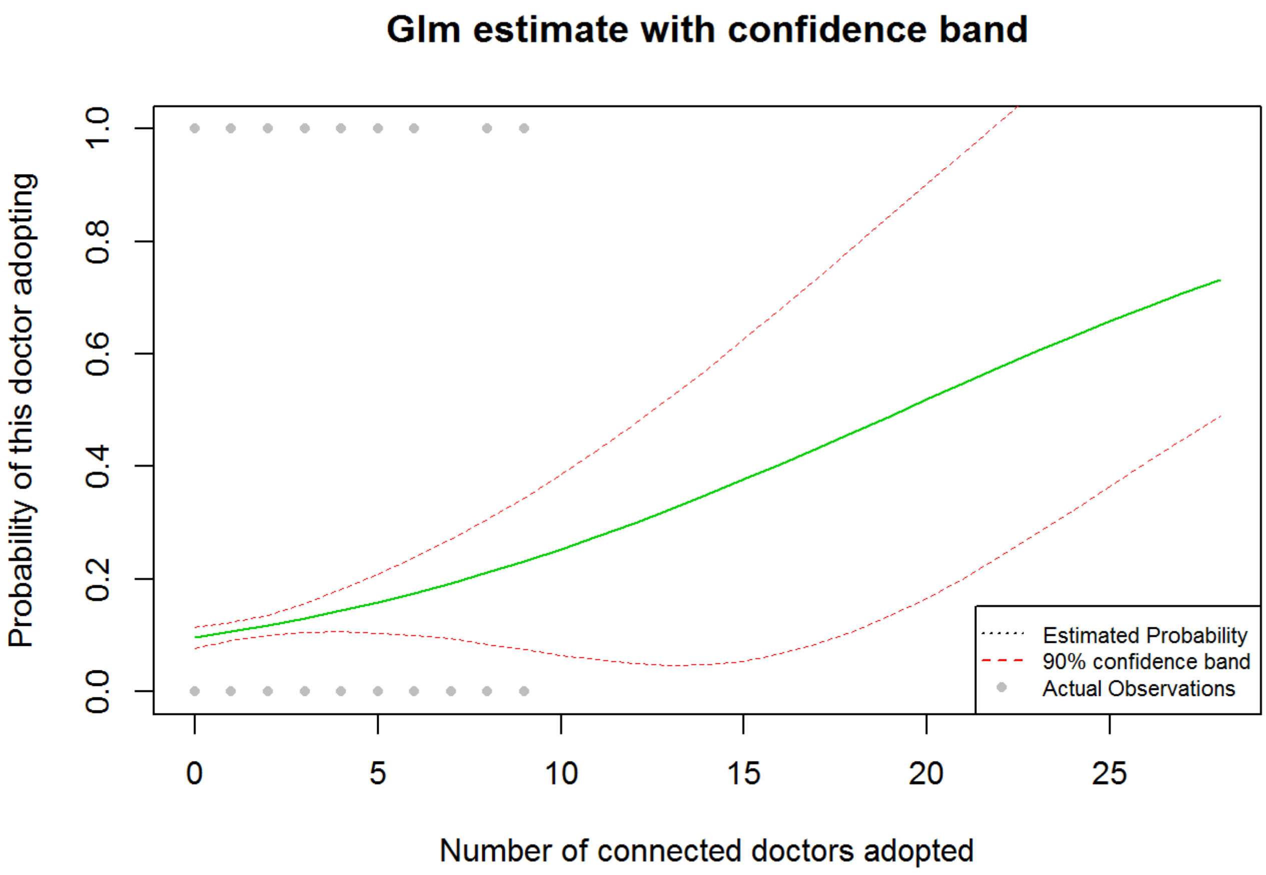


Given our model, we want to see how the probability of a doctor prescribing changes with the total number of doctors prescribing. We want to discover if there is any underlying relationship between the two variables. We plot the estimated probabilities across all possible number of doctors adopting the drug in our model, and we observed that the probabilities was highest when the total number of doctors adopted being around 60, and are lower on both ends of the spectrum. This is not sufficient for us to make a conclusion with regard to the relationship between the probability of a doctor adopting the drug and the total number of doctors already prescribing. Also, since the probability and the total number of doctors adopting changes every month and each value of total number of doctors adopting is applied to a certain set of doctors, we want to take the average predictive comparison across all doctors and months to determine the effect of number of change in doctors adopting and the probability of a new doctor adopting.

We used bootstrap resampled data to compute the average predictive comparison averaging over doctors and months for our given model. This method with uncertainty provides a more accurate estimation of the true value and thus make our prediction and conclusion more accurate. Given that the average predictive comparison for our kernel model is -2.6110^{-6} and a standard error of 1.3710^{-20} , we conclude that with one unit of change in total number of doctors prescribing, the probability of a doctor adopting the drug at the given month is expected to change by -2.6110^{-6}. We conclude that the relationship between the probability and the total number of doctors is rather insignificant.

We then proceed to check the potential relationship between the number of doctors in the doctor’s network adopting the drug and the probability of this doctor adopting the drug at a given month. Given the information we know about each doctor and their network, we constructed a data frame that contains all combinations of doctors and months, and their status in terms of whether they are adopting the drug this month, and whether this doctor has already been prescribing the drug, and the total number of other doctors in his network prescribing the drug in this given month. Given the assumption that all probabilities across all months of doctors prescribing are the same, we want to find out the relationship between the probability of a doctor begin to adopt the drug at a given month and the total number of doctors in his network already prescribing.

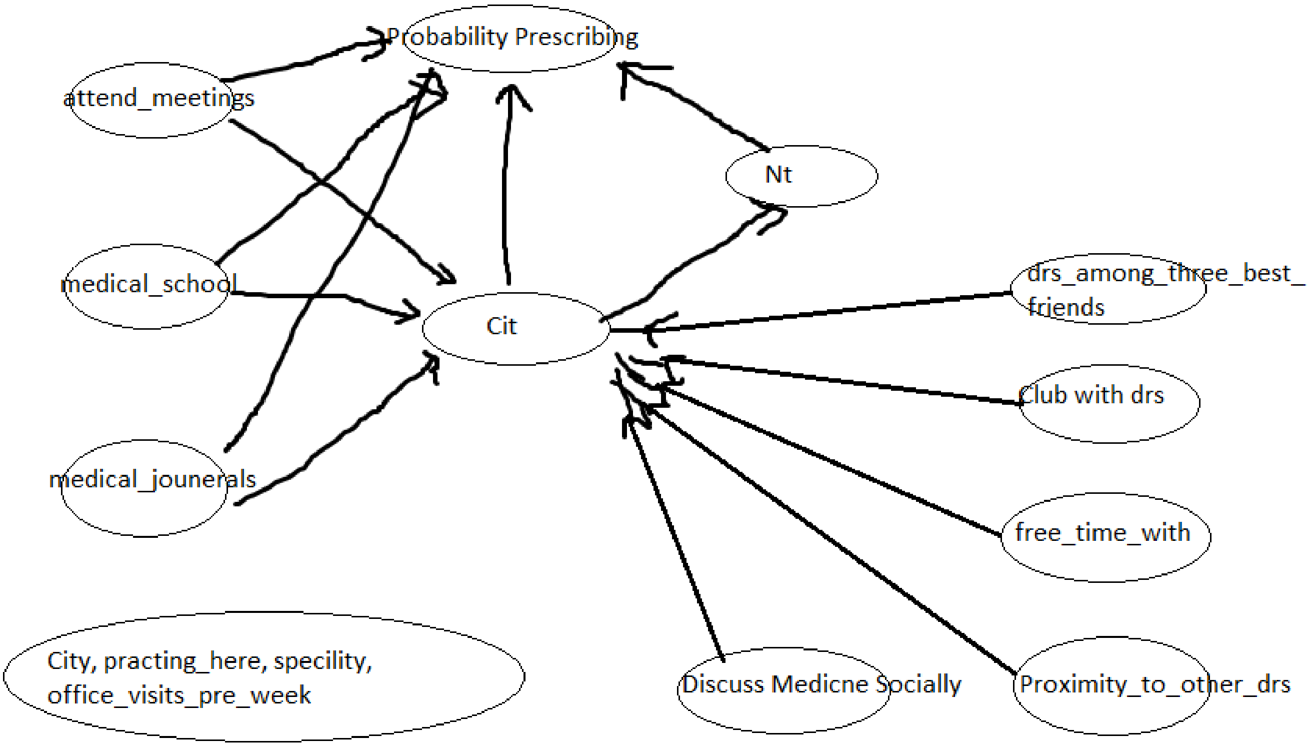
We decided to use generalized linear models with logistic regression to make our predictions. Even if it does not provide as accurate of a prediction as our kernel regression, it can make predictions beyond our given range of number of doctors in network prescribing (x-values). Kernel preforms poorly as its prediction turns flat when there are no observations at the extended range. Since we have a very limited number of x-values (from 0 to 9) and we want to make a prediction at a longer range of x-values, we decided to move on with generalized linear model.



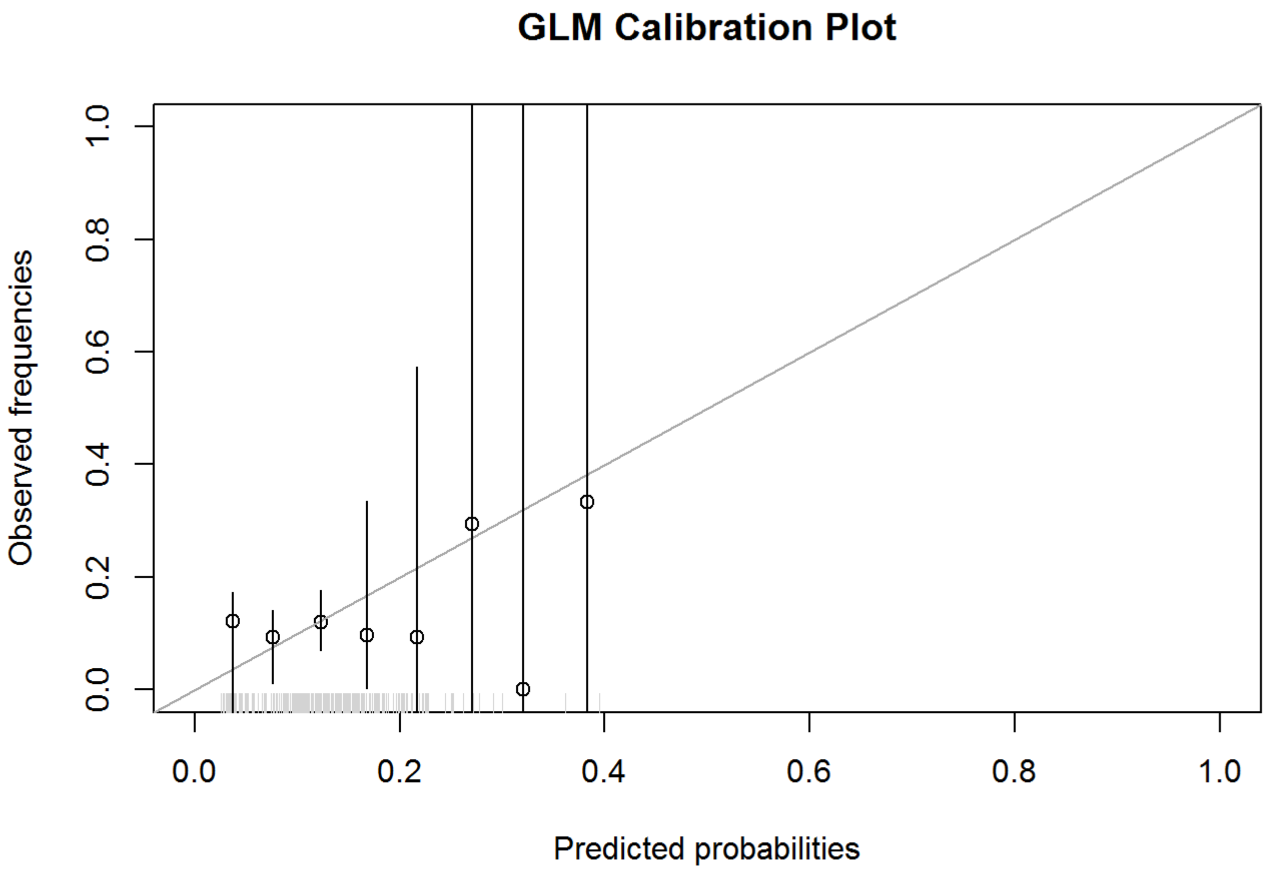
We produced the estimated probability of a doctor adopting the drug at given month t given the total number of doctors prescribing in his network. We choose the range from 0 to 28 because the maximum number of connection of any given doctor in our entire network matrix is 28. We wish not to extend our prediction beyond this maximum possible range. Even though our GLM model preforms poorly beyond certain range as the bootstrapped confidence band becomes more apart from the prediction line, we can conclude that there is a higher number of people prescribing in the network of a given doctor is related to the probability of this doctor adopting this drug at a given month. We further confirm our statement with average predictive comparison. Averaging across all doctors and months, we have the bootstrapped mean average predictive comparison 0.116 with standard error of 1.2610^{-17}. This means there with an extra number of doctors prescribing the drug in a doctor's network, he or she is more likely to adopt the drug by a probability of 0.116.

Determine casual relationships

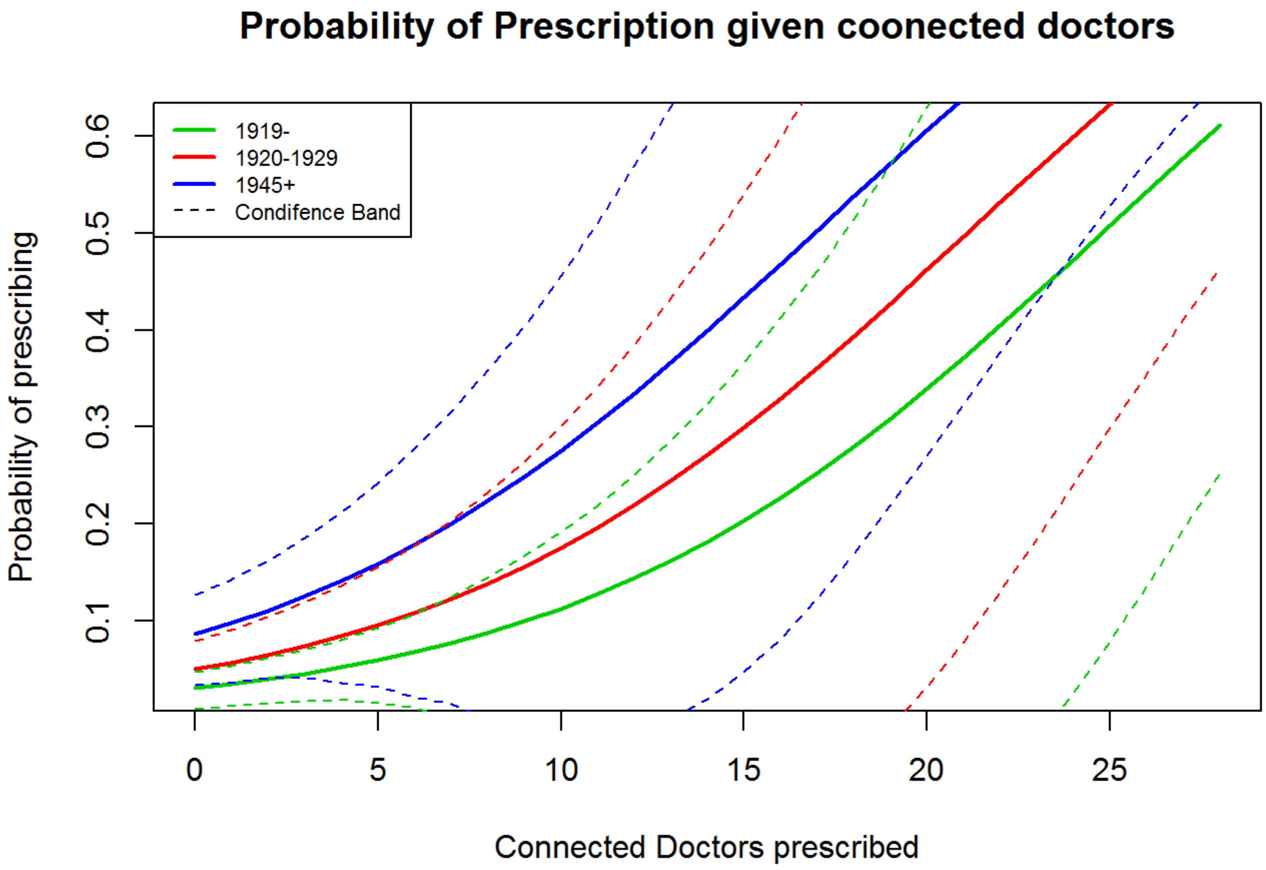
We conclude from our previous result that the probability of a doctor adopt the new drug at given month t is not related to the total number of doctors prescribing the drug at that given month. However, such probability is related to the number of doctors that have already adopted the drug at the given month. Our findings are not consistent with one and another because there may be potential causal relationship between doctors within the network prescribing (Cit) and total number of doctors prescribing (Nt) and the probability of a doctor prescribing. Even though the total number of doctors prescribing in the network is a subset of the total number of doctors prescribing, we deduce that there is a causal relationship between the number of doctors in network prescribing and the total number of doctor prescribing. In particular, total number of doctors prescribing and the probability of a doctor prescribing shares a common ancestor, which is the doctors in network prescribing.



We also want to explore other casual relationships. We believe that there is casual relationship between the meetings that the doctors attend, when the doctor finished his or her medical school, and how many medical journal a given doctor reads over the period of time, and probability of this doctor prescribing, and the probability of this doctor's total number in the network adopting. We also believe that there are potential causal relationship between proximity to other doctors, who the doctor spends his free time with, if the doctor is in the same club as other doctors, if the doctor discuss medicine socially, and whether the doctors have other doctors amongst his or her three best friends. They all are measure of friendships and they all have a casual effect on the doctors' total number of network prescribing.



Therefore, we want to find out the probability of a doctor adopt the new drug at a given month and the total number of doctors adopted the drug in his or her network, control for his medical society meeting attendance, number of journals read, and the year he graduated from medical school. Particularly, we are interested in the casual effect of year graduating from medical school on the probability of doctor adopting the drug. We decided to use a general linear model with logistic link for our model. Similar to previous reasons, since we have a small amount and one sided x-values and we want to predict over a wider range of values, we do not think non-parametric regression is appropriate in this case because they would need more data down the line to make better prediction. From examining the frequency vs probability plot of GLM and its prediction, we conclude that the GLM model have an appropriate goodness of fit. However, it does not predict well over higher probabilities.



We continue to examine the effect of med school graduation year on the relationship between probability of a doctor adopting the drug and the total number of doctor in his network adopting the drug, control for medical society meetings and number of journal read. We can observe from graph that as the doctors graduate from medical school later, they are more likely to adopt the drug than those who graduated earlier (as they horizontally shifts upwards). Also, from each curve we observed form the figure above, we deduce that when control the graduation year of the doctors, the more people in the doctor's network is prescribing the drug, the more likely the doctor is going to adopt the drug at this given month. We continue testing this with the average predictive comparison of our model. As the number of doctors in this given doctor's network increase by one, the probability of this doctor adopting the drug in a given month increases by 0.14 with standard error of 1.3410^{-16}, under the assumption that all other backdoor paths to the probability of adoption and total number of adoption in network is blocked. Meaning that all other causal effects are controlled for in this model, and there does not exist a hidden causal relationship in our model.

Conclusion

From our observation and modeling of our data, we reached the conclusion that the probability of a doctor adopting the new innovative drug in a given month has a causal relationship with the total number of doctors prescribing in his network, the year he graduated from medical school, whether the doctor attends medical society meetings, and the number of medical journal that he or she reads. We concluded that the more people in doctor's network are prescribing the drug, the more likely this doctor is going to prescribe the drug at any given month. Also, we found out that the more recent the doctor graduates from medical school the more likely the doctor is to adopt the new antibiotics. Surprisingly, the probability of a doctor adopting the drug is not increased with the total number of doctors adopting the drug, reinforcing the point that ideas are passed along one and another through personal contact. While we identified the direct causal relationship between the probability of adopting the drug, total number of doctors in network adopting the drug, year of med school graduation, medical society meeting attendance and frequency of reading medical journals, we also believe that there exists causal relationship between total numbers of doctors in network prescribing. In particular, we believe that a doctor's proximity to other doctors, whether he spends his free time with doctors, whether the doctors' best friends are also doctors, whether the doctor is in a club with another doctor, and whether the doctor discuss medicine socially have a causal relationship with the total number of doctors in his or her network prescribing drugs. We consider those doctors who spends his free time around with other doctors are more likely to know more doctors, and thus are more likely to be exposed to the doctors that are prescribing the new drug.