Exam 3: How Tetracycline Came to Peoria

36-402, Advanced Data Analysis

Due at **10:30 am** on Monday, 11 May 2015

> *Thoughts, like fleas, jump from man to man. But they don't bite everybody.* — Stanislaw Lem

Now-common ideas like "early adopters" and "viral marketing" grew from sociological studies of the diffusion of innovations. One of the most famous of these studies tracked how a then-new antibiotic, tetracycline, spread among doctors in four small cities in Illionis in the 1950s. In this exam, we will go back to that data to look at one of the crucial ideas, that of the innovation (prescribing tetracycline) spreading from person to person through personal contact.

On the class website, you will find two data files, ckm-nodes.csv and ckm-net.dat. The former has information about each individual doctor in the four towns. `adoption_date` records the month in which the doctor began[1] prescribing tetracycline, counting from November 1953. If the doctor did not begin prescribing it by month 17, i.e., February 1955, when the study ended, this is recorded as `Inf`. If it's not known when or if a doctor adopted tetracycline, their value is `NA`. Other columns record when the doctor graduated from medical school, whether they attend medical conferences recently (and if so, what kind), how many medical journals they read, and other information about the individual doctors. Note that the covariates in this file are a mix of ordinal variables, categorical variables, and numerical variables.

The `ckm_network.dat` file contains a binary matrix, which records the social network among the doctors. There is one row and one column for each doctor; the i, j entry is 1 if doctor number i and doctor number j knew each other, and `0` if they did not.

1. (5) Create a plot of the number of doctors who *began* prescribing tetracycline each month versus time. (It is OK for the numbers on the horizontal axis to just be integers rather than formatted dates.) Produce another plot of the *total* number of doctors prescribing tetracycline in each month. (The curve for total adoptions should first rise rapidly and then level out around month 6.)

2. Estimate the probability that a doctor who had not yet adopted the drug will begin to do so in a given month t, as a function of the *total* number of doctors N_t who had adopted *before* t. (You may assume that these probabilities are the

[1] Apparently, no one started prescribing the drug and then stopped.

same for all t.) You may estimate this function however you like, but be sure to explain how you are estimating these probabilities, and how you know that method is reliable *in this particular case*. (This may involve model checking.)

(a) (5) Report these probabilities as a curve, with N ranging from 0 to 125. If you do not think you can estimate the whole range, plot as much as you can, and explain why you cannot go further. For full credit, your plot must have more than 17 points. Also for full credit, your curve should be accompanied by some measure of its error.

(b) (5) Averaging over doctors and months, how much does the predicted probability of adoption change when N increases by 1? Give a standard error to this change in predicted probabilities.

Hint: You might try building a new data frame which records, for each month, the number of doctors who adopted tetracycline that month, and the number who had previously adopted tetracycline.

3. Estimate the probability that a doctor i who had not yet adopted the drug will begin to do so in month t, as a function of the number C_{it} of *doctors linked to* i who had adopted before t. (Again, you may assume that these probabilities are the same for all t.)

(a) (8) Make a plot of these probabilities, with C_{it} ranging from 0 to 30. If you do not think you can estimate the whole range, plot as much as you can, and explain why you cannot go further. For full credit, your plot must include at least 29 points, and include a measure of uncertainty in your estimates. Does your curve support the idea that the use of tetracycline is transmitted from one doctor to another through the social network? Explain, including a description of what curves which did *not* support this idea would look like, or why the shape of this curve is actually irrelevant to this issue.

(b) (7) Averaging over doctors and months, how much does the predicted probability of adoption change when C_{it} increases by one? What is your standard error for this change in predicted probabilities?

Hint: You might try building a data frame recording, for every combination of doctor and month, whether that doctor began prescribing tetracycline that month, whether that doctor has begun prescribing tetracycline earlier than that month, and the number of their contacts who began prescribing before that month.

4. (a) (1) Are your estimates from problem 2b and 3b consistent with one another? Explain.

(b) (4) What would you have to assume for either of these to be estimates of the causal effect on adoption by other doctors of making one extra doctor adopt the drug? Be as specific as you can, rather than just repeating definitions from the notes. Drawing graphs is encouraged.

5. Estimate a model which predicts the probability that a doctor i who had not yet adopted the drug by month t will begin to do so in month t, as a function of C_{it} and of the covariates which indicate when i went to medical school, whether they attended medical-society meetings (and if so what kind), and how many medical journals they read.

   (a) (5) Plot the estimated probability of adoption as a function of C_{it} for doctors who read the minimal number of journals, do not attend conferences, and graduated from medical school (i) in 1919 or earlier, (ii) in the 1920s, and (iii) in 1945 or after. For full credit, have all three lines on the same plot (clearly visually distinct from each other), and some measure of uncertainty for each line.

   (b) (5) Averaging over doctors and months, how much does increasing C_{it} by one change the probability of doctor i adopting tetracycline in month t? Include a standard error for this change in predicted probabilities.

   (c) (5) Under what assumptions does this give a valid estimate of the average causal effect of increasing C_{it} by one?

1 Formatting Instructions and Rubric

Your main report should be a humanly-readable document of at most 10 single-spaced pages, including figures. It should have the following sections:

INTRODUCTION describing the scientific problem and the data set, possibly including *relevant* summary statistics or exploratory graphs. (Do not include EDA just to have EDA.)

SPECIFIC PROBLEMS answering the questions set above, but avoiding the check-list, itemized format in favor of continuous text, with a logical succession of sentences and paragraphs. (Writing coherently is more important than following the order of the questions.)

CONCLUSIONS summarizing what you have learned from the data and models about whether the transmission of an innovation from person to person is really a good description of how these doctors came to use tetracycline.

You may assume that the reader has a general familiarity with the contents of 401, and with the models and methods we have covered so far in the course, but will need to be reminded of any details. The reader should not be assumed to have any prior familiarity with the data set.

Code *All* statistical results must be supported by appropriate code, or they will receive no credit. ("Show your work.") Code should only appear in the text of the report when it is the best way of conveying some point. The ideal would be to use R Markdown, or knitr+LaTeX, to embed all computations in a humanly readable document, and submit both the knitted version and the source[2] As a second best, it is acceptable to submit a PDF document containing all text and figures, and a separate .R file, containing all supporting computations, clearly labeled via the comments so that it is easy to see which claims or results go with which pieces of code.

Exploratory data analysis Do as much EDA as you like. (If you want to display the social network, the R package `igraph` is designed for such things.) Do not include any of this EDA in your report, expect for plots, tables, etc., which *specifically* matter for either your methods of analysis or your conclusions, in which case, explain their relevance. No credit will be given for unmotivated EDA, no matter how elaborate.

Rubric

As usual, this describes the ideal.

Words (5) The text is laid out cleanly, with clear divisions and transitions between sections and sub-sections. The writing itself is well-organized, free of grammatical and other mechanical errors, divided into complete sentences logically grouped into paragraphs and sections, and easy to follow from the presumed level of knowledge.

Numbers (5) All numerical results or summaries are reported to suitable precision, and with appropriate measures of uncertainty attached when applicable.

Pictures (5) Figures and tables are easy to read, with informative captions, axis labels and legends, and are placed near the relevant pieces of text.

Code (15) The code is formatted and organized so that it is easy for others to read and understand. It is indented, commented, and uses meaningful names. It only includes computations which are actually needed to answer the analytical questions, and avoids redundancy. Code borrowed from the notes, from books, or from resources found online is explicitly acknowledged and sourced in the comments. Functions or procedures not directly taken from the notes have accompanying tests which check whether the code does what it is supposed to. All code runs, and the Markdown file knits (if applicable). The main text of the report is free of intrusive blocks of code, which are used only when a specifically-computational point is being made, or when code is actually the clearest way of describing a point.

Inference and Uncertainty (10) The actual estimation of model parameters or estimated functions is technically correct. All calculations based on estimates are clearly explained, and also technically correct. All estimates or derived quantities are accompanied with appropriate measures of uncertainty (such as confidence intervals or standard errors). Default standard errors, confidence intervals, and other inferential statistics produced by R or its packages are used only when there are specific reasons

[2] See examples at `http://yihui.name/knitr/demos/`, and the useful chunk options like `echo` at `http://yihui.name/knitr/options/`; also the examples in the solutions to exam 1.

to think they are valid for the present purpose. (Partial credit may be given if they are used because you cannot get anything else to work.)

Conclusions (10) The substantive questions about diffusion of innovations are all answered as precisely as the data and the model allow. The chain of reasoning from estimation results about models, or derived quantities, to substantive conclusions is both clear and convincing. Contingent answers ("if X, then Y, but if Z, then W") are likewise described as warranted by the model and data. If uncertainties in the data and model mean the answers to some questions must be imprecise, this too is reflected in the conclusions.

Extra credit (10) Up to ten points may be awarded for reports which are unusually well-written, where the code is unusually elegant, where the analytical methods are unusually insightful, or where the analysis goes beyond the required set of analytical questions. *Example:* Simulating the model estimated in problem 5, taking the set of doctors who have adopted in month 1 for the initial conditions and continuing for another 16 months, with a detailed and quantitative comparison of multiple simulation runs to the actual data, and an informative assessment of what the comparison says about the strengths and weaknesses of the model.