Using Quantitative Methods in Industry

Olivia Lau, *Google, Inc.*
Ian Yohai, *Netflix, Inc.*

While we both focused on quantitative methods as graduate students, we do not recall anyone who expressed a wish to become a data scientist, even among those seriously considering careers outside of academia. We completed our PhDs in 2008 and 2009 respectively, and have since been fortunate to spend time at the Food and Drug Administration; as consultants for the US Agency for International Development, the Department of Defense, the Defense Advanced Research Projects Agency, and other federal agencies; and in large technology companies in Silicon Valley.

Having followed the data science path for several years, we have identified some areas where political science quantitative methods training overlaps with the essential elements of the data scientist's toolkit: experimental design and causal inference; design and analysis of survey data; predictive modeling; and finally implementing methods in computing languages such as R or Python. In all the areas we discuss, one's employability increases markedly as the level of technical depth increases.

EXPERIMENTAL DESIGN AND CAUSAL INFERENCEE

Experimental design and causal inference are extremely important in the industries in which we have worked, particularly in pharmaceuticals and technology companies. In contrast to the observational nature of many social science analyses, experiments in the industries in which we have worked often start with a randomized design.

For internet-based technology companies, it is relatively easy to run experiments over a sample of users. Hal Varian, Chief Economist at Google, points out that 'kaizen', a philosophy of continuous incremental improvement through iterative experimentation, is the heart of engineering innovation (Varian 2007). Engineers routinely run randomized experiments to determine if a change has an effect on targeted outcomes (c.f. Kohavi, Henne and Sommerfield 2007).

Likewise, in pharmaceuticals, experimental design is a critical component of assembling clinical trials protocols for regulatory submission, review, and approval. Before a drug or medical device trial starts, representatives from the pharmaceutical company and the relevant regulatory agency agree to a long list of issues, including: the relevant patient population, the method of selecting trial participants, the randomization of participants to test and control, the clinical outcomes to be recorded at agreed-upon intervals, the margin between treatment and control groups that would constitute a successful trial, and sometimes even the specific statistical test that would be used to calculate the trial's success or failure.

Data scientists should have the ability to develop solid research designs and test assumptions to ensure that the inferences being drawn are reliable. When randomized experiments are appropriately designed and executed, they may be analyzed with relatively straightforward methods, such as a simple difference in means, test of equal proportions, or a chi-square test. When experiments involve testing three or more treatments simultaneously, the data scientist also needs to understand interaction effects and proper factorial design.

Running randomized experiments in the real world often involves practical challenges that can significantly affect inferences. Consider a few examples: clinical trial participants may not take the prescribed dosage at the prescribed time, which means that patient outcomes may or may not be a result of the drug or device under study. In industry, many marketing applications pose a challenge since exposure to some types of advertising like TV and radio cannot be controlled at the individual-level. In other instances, if users assigned to control find out about the new feature being tested in the treatment group, the experiment might be contaminated if the control units try to game the system to access the new feature.

Even if the applications are somewhat different, the skills learned in political science graduate programs are directly relevant to the analysis of experimental data in industry. Common methods employed in the analysis of imperfect experiments include propensity score matching, instrumental variables estimation, and causal graphical models, methods similar to those used in political science for the analysis of field experiments, natural experiments, and observational data.

In addition, quasi-experimental methods may be employed to help make causal inferences when fully randomized designs are not possible. These quasi-experiments may take the form of region-based comparisons, such as when an intervention is done in one set of countries or regions but not others. A sound grounding in time series analysis and related regression techniques is required since many studies need to compare a pre-period without any intervention to a post-period after the intervention has occurred (c.f., Lambert and Pregibon 2008). Simply running a regression is not usually enough; but rather being able to exploit natural or induced variation in a clever way can dramatically improve the study design.

doi:10.1017/S1049096516000901

We believe this is a good example where seeking greater technical 'depth' as part of a graduate program will help candidates stand out during the interview process. Many advanced statistics courses on causal inference cover methods for analyzing quasi-experiments and addressing non-compliance, but these should be more regularly offered as part of political science programs.

SURVEY DESIGN AND ANALYSIS

Given that survey analysis has deep roots in political science, there is a good fit between the training provided in many graduate programs and the requirements for being a survey analyst in industry. Surveys are particularly useful for identification of trends and market sizing.

Companies frequently conduct surveys to provide insights into market and consumer trends. In many cases, surveys complement experimentation since a test cannot necessarily explain why a particular product feature was successful or unsuccessful. Conducting surveys in conjunction with experiments can sometimes provide a story to fill this gap. When companies operate in distinct markets, surveys may be useful in understanding cross-cultural and market-specific features that may help or handicap a product launch. Surveys can also help inform product deployment decisions by estimating the demand for a new product. For example, epidemiological studies can inform drug development decisions by estimating the prevalence of specific diseases in the population (c.f., Parekh et al. 2011).

Surveys in industry face the same challenges as in other areas such as political and opinion polling, particularly low and declining response rates. Accordingly, good survey analysts will have familiarity with statistical techniques like non-response weighting and missing data imputation. There are also a variety of modes in which surveys are conducted: in-product, e-mail, Internet panels, and occasionally by telephone. Research on mode effects is thus highly relevant.

PREDICTIVE MODELING

Predictive modeling has become crucial to many companies with big data problems. Access to huge datastores is not enough by itself without methods to derive actionable insights. From recommendations systems that give personalized viewing suggestions to subscribers to detecting potential fraud to modeling when customers are likely to cancel subscriptions, good predictive models can materially affect the bottom line. Both parametric modeling and machine learning methods may be employed to developing these models.

In our experience, most political science graduate programs emphasize parametric modelling, with less attention paid to unsupervised machine learning. Certainly, having a solid grounding in linear regression and related methods is a prerequisite for many positions that involve predictive modeling. In addition, survival models that predict time to failure are frequently used to model the rate at which customers are likely to cancel service (Van den Poel and Larivière, 2004).

Exposure to machine learning methods, such as support vector machines, random forests, boosting techniques, and other methodologies is often required for many positions. Some roles also necessitate familiarity with natural language processing (NLP) methods, including topic modeling and sentiment analysis. In the last several years, these methods have also gained popularity in political science (Hopkins and King 2010; Quinn et al. 2010), so graduate students may be familiar with them. However, many roles require substantial depth in machine learning methods, and so candidates considering a career in industry may want to supplement the traditional graduate coursework with additional training. As we discuss below, knowledge of programming languages is also essential, and this is particularly true in the machine learning area as training and validating models requires substantial data processing.

Common methods employed in the analysis of imperfect experiments include propensity score matching, instrumental variables estimation, and causal graphical models, methods similar to those used in political science for the analysis of field experiments, natural experiments, and observational data.

COMPUTING AND VISUALIZATION

We were fortunate that in our graduate program, R was the required language in methods courses; we both became fairly proficient R programmers. This was helpful not only because R is widely used in industry, but also because facility with one programming language makes it somewhat easier to learn another and also signals to employers that the applicant can do so.

Proficiency with R (or Python) is not the only requirement for aspiring data scientists. For our research in graduate school, we worked with datasets on the order of hundreds of thousands of rows. Now we routinely work with datasets on the order of billions of rows. SQL is therefore widely used to pull data. While neither of us learned SQL in graduate school, understanding equivalent concepts in R (e.g., "merge" being similar to "join") was helpful. Nevertheless, more focus on databases would likely be beneficial in graduate programs, and would also enable more researchers to undertake large scale analyses in political science as well.

Data visualization has also taken on increased importance, since many companies track a slew of metrics on a daily basis or even in real-time. These must be intelligently presented, however, since it is easy to lose track of what is most important when so much data are available. Many data scientists

develop dashboards that concisely summarize information and that allow the user to draw his or her own conclusions. This in turn requires both technical skill in creating the visualizations and the analytic insight to highlight that which is most critical. Fortunately, tools like Shiny are now available that enable interactive visualizations within web applications that are easily shared across users. Someone already familiar with R can thus create web apps without needing to learn JavaScript. These tools should be included in methods courses so that graduate students can gain all the skills necessary to be a successful data scientist.

For our research in graduate school, we worked with datasets on the order of hundreds of thousands of rows. Now we routinely work with datasets on the order of billions of rows. SQL is therefore widely used to pull data.

CONCLUSION

Graduate students who take quantitative methods courses in political science will be prepared to enter the workforce as data scientists. Experimental design, causal inference, survey design and analysis, and predictive modeling are all highly relevant in today's big data environment. Candidates considering such a path may wish to go beyond the standard coursework, especially in the areas discussed above. Some examples include courses that provide a rigorous grounding in causal inference (both for experimental and quasi-experimental designs); advanced survey design and analysis; advanced statistical modeling (such as survival modeling), and supervised and unsupervised machine learning. Graduate programs should also evolve their curriculum to keep pace with the tools (such as SQL, R, and Python) that are used in industry. Finally, graduate students should engage in professional development by presenting research at conferences and other forums. The communication skills learned in graduate school are equally as important as the technical skills to be successful in today's workforce. ■

REFERENCES

Hopkins, Daniel and Gary King. 2010. "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science* 54 (1): 229–247.

Kohavi, Ron, Randall M. Henne, and Dan Sommerfield. 2007. "Practical Guide to Controlled Experiments on the Web: Listen to Your Customers not the HiPPO." KDD'07, August 12–15. San Jose, California.

Lambert, Diane and Daryl Pregibon. 2008. Online Effects of Offline Ads. In ADKDD '08: Proceedings of the 2nd International Workshop on Data Mining and Audience Intelligence for Advertising, ACM: New York, 10–17.

Parekh, Anand K., Richard A. Goodman, Catherine Gordon, Howard K. Koh, and The HHS Interagency Workgroup on Multiple Chronic Conditions. 2011. "Managing Multiple Chronic Conditions: A Strategic Framework for Improving Health Outcomes and Quality of Life." *Public Health Reports* 126 (4): 460–471.

Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54 (1): 209–228.

Van den Poel, Dirk and Bart Larivière. 2004. "Customer Attrition Analysis for Financial Services Using Proportional Hazard Models" *European Journal of Operational Research* 157 (1): 196–217.

Varian, Hal. 2007. Kaizen, That Continuous Improvement Strategy, Finds Its Ideal Environment. *The New York Times*, February 8. http://www.nytimes.com/2007/02/08/business/08scene.html?_r=0