METHODOLOGY

Can Biomedical Researchers Benefit from Political Science Methodologists?

JOSUÉ GUZMÁN, PhD

The purpose of this monograph is to introduce the work of a group of political science methodologists, who created a series of models for estimation and prediction. It refers to the work of Gary King and his collaborators at Harvard University's Department of Government. They developed a set of computer-intensive simulation techniques, along with the corresponding public domain software. Their paper, "*Making the most of statistical analyses: improving interpretation and presentation.*" and the corresponding software, *Clarify* (Tomz *et al.*, 1999) are the basic sources. The use of statistical simulation has become a standard practice with the advent of fast and low cost modern computing. The reader can appreciate the usefulness in biomedical research of *Clarify*. It can be of much help on the statistical analysis of many biomedical problems.

Key words: Biomedical research, Models, Monte Carlo simulation, Estimation, Prediction, Clarify, Stata.

What is the relationship between a presidential election vote and the probability of prostate cancer? Does the Mexican presidential election have anything to say about suicide characteristics in another country? What about political leader tenure and a patient's recurrence after treatment? The initial answer to the above questions is that there is no relationship. However, political science experts are nowadays using modern statistical methods that, if adopted, could benefit biomedical researchers.

These new statistical methods, instead of being based on theory and mathematical calculus, are using computer-intensive techniques. Moreover, instead of using classical significance hypothesis testing, are relying more on estimation of characteristics of interest.

The purpose of this monograph is to introduce the work of a group of political science methodologists, who created a series of simple but powerful models for estimation and prediction. Specifically, we refer to the work of Gary King and his collaborators at Harvard University's Department of Government (http://GKing.Harvard.edu). They developed a set of computer-intensive simulation techniques, along with the corresponding public domain software. We cite their paper, "*Making the most of statistical analyses: improving interpretation and presentation.*" (1) and the corresponding Stata macros known as *Clarify*© (2).

King and collaborators introduce user-friendly techniques, "to convert raw results ... into expressions that (i) convey numerically precise estimates of the quantities of ... interest, (ii) include reasonable measures of uncertainty about those estimates, and (iii) require little specialized knowledge to understand." (1) Furthermore, they suggest that the approach:

- "...can extract new quantities of interest from standard statistical models ...
- ... allow scholars to assess the uncertainty surrounding any quantity of interest ..., and
- can convert raw statistical results into results that everyone ... can comprehend."

The King and Collaborators' Methodology

Their methodology relies on a general class of two-equation statistical models. Let X be a set of explanatory variables (or covariates) to a characteristic of interest (or response) y. The King and collaborators' model have two filters that link X to y: a *systematic component*, $\theta=g(X,\beta)$

From the Department of Biostatistics and Epidemiology, Graduate School of Public Health, Medical Sciences Campus, University of Puerto Rico, San Juan PR.

This study was supported by the RCMI Program of the University of Puerto Rico, Medical Sciences Campus (NIH G12-RR-03051). A travel grant to attend the 2000 Summer Session on Public Health Studies at the Harvard School of Public Health.

Address correspondence to: Dr. Josué Guzmán, Department of Biostatistics and Epidemiology, Graduate School of Public Health, Medical Sciences Campus, University of Puerto Rico, San Juan PR 00936-5067. Tel. (787) 758 2525 Ext. 1428, Fax: (787) 764 5831, E-mail: jguzman@rcm.upr.edu

that depend on the values of the explanatory variables, X, and the effect parameter β, and a *stochastic component*, $f(\theta,\varphi)$ a probability density that generates the characteristic of interest, y, where φ is a possible ancillary parameter. This can be represented as (we are using a slightly different notation):

$$g(X,\beta) \rightarrow f(\theta,\varphi) \rightarrow y.$$

An illustration of such model is where y is dichotomous (patient recurrence, *yes* or *no*) and a *logistic* regression model is used to express its relationship with X:

$$g(X,\beta) = \exp(X\beta).[1+\exp(X\beta)]^{-1}$$
$$f(\theta,\varphi) \sim Bernoulli(\pi) \rightarrow y,\ \pi = p(\text{recurrence})$$

The logit and other models are standard techniques to biomedical researchers. The novelty of the King and his collaborators' approach is that they distinguish between two different kinds of uncertainty: *fundamental* (or model specification) uncertainty and *estimation* uncertainty. Fundamental uncertainty is due to the fact that there are an infinite number of explanatory variables, apart from X, that could affect y, but are not included in the model. Estimation uncertainty arises due to the lack of knowledge of φ and β; thus, their estimates, $\hat{\varphi}$ and $\hat{\beta}$ are uncertain.

Via computer-intensive techniques, one can simulate these uncertainties in order to get results, which are easier to interpret than statistical significance testing results (*p-values*, α levels). The technique is known as *Monte Carlo* (*MC*) simulation. *MC* simulation generates M (pseudo) random samples from a probability distribution, $p(y)$. From the M draws, we can calculate any function of y along with a $(1-\alpha)$% interval, by just taking the $\alpha/2$ and $1-\alpha/2$ values of the ordered observations of the simulation.

Monte Carlo Simulation

How this is done with the King and collaborators' approach? They start by considering a column vector

$$\hat{\gamma} = \begin{pmatrix} \hat{\beta} \\ \hat{\varphi} \end{pmatrix}$$

with associated variance matrix $\hat{V}(\hat{\gamma})$. Based on the *Central Limit Theorem*, form a multivariate normal distribution $\gamma \sim N[\hat{\gamma}, \hat{V}(\hat{\gamma})]$. Then, simulation is done via the following algorithm:

1. Estimate the model, and obtain $\hat{\gamma}$ and $\hat{V}(\hat{\gamma})$
2. Draw one value of γ from the above multivariate normal model, denote it
$$\tilde{\gamma} = \begin{pmatrix} \tilde{\beta} \\ \tilde{\varphi} \end{pmatrix}$$
3. Repeat steps 1 and 2 M times.

The above algorithm forms the basis to obtain mean (or expected) values, difference between two mean values (first differences), and predicted values (1).

The King and collaborators' approach can be implemented by a set of three macros written for the *Stata*™ statistical program, known as *Clarify* (2):

- *estsimp* – to estimate the specified model and to generate samples to calculate (default M=1,000 samples).
- *setx* – to specify the values of the explanatory variables, X.
- *simqi* – to compute the desired quantities of interest: mean values, first differences and predicted values.

Biomedical Illustrations

1. Predicting nodal involvement in prostate cancer patients. The purpose of a study on prostate cancer patients, reported in (3) was to determine which combination of five prognostic factors could be used to forecast whether or not the cancer has spread to the surrounding lymph nodes. The data have also been analyzed in (4) and in (5). The response variable is nodal involvement [*ni*] (0 = absence, 1 = presence of nodal involvement). Potential prognostic factors are: level of serum acid phosphatase [*ap*] and age as variates, and X-ray result (0 = negative, 1 = positive), tumor size or stage (0 = small, 1 = large), and tumor pathological grade (0 = less serious, 1 = more serious) as factors. Using the *parsimony principle* (among competing models, choose the simplest, and informative, one), we found that the level of acid phosphatase (log transformed [*lap*]) and the X-ray were the best predicting variables. Furthermore, the logit model (without the constant term) appears to be adequate to explain the probability (π) of nodal involvement. That is,

$$\text{logit}(\pi) = \beta_1 \cdot \log(ap) + \beta_2 \cdot xray,$$

where logit (π) = $\log[\pi/(1-\pi)]$, log (ap) = log(acid phosphatase), $xray$ = X-ray result.

Using *Stata*, we estimated (by maximum likelihood estimation) the pertinent model:

```
. logit ni lap xray, nocon nolog

Logit estimates                                   Number of obs = 53
Log likelihood = -30.046681
```

ni	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lap	-.2647103	.0902093	-2.934	0.003	-.4415173	-.0879033
xray	2.145047	.7006817	3.061	0.002	.7717361	3.518358

Thus, our estimated model is

$$\text{logit}(\pi) = -.265 \cdot \log(ap) + 2.145 \cdot xray$$

Using *Clarify*'s ado files within *Stata*, we simulated the model ($M = 1{,}000$ samples) with the odds ratio option:

```
. estsimp logit ni lap xray, nocon or nolog
Logit estimates                                   Number of obs =  53
Log likelihood = -30.046681
```

ni	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lap	.7674283	.0692292	-2.934	0.003	.64306	.9158495
xray	8.542442	5.985532	3.061	0.002	2.163519	33.72899

We observe an adjusted odds ratio of 0.77 for each unit of increase in log acid level (95% CI: 0.64 - 0.92), and an adjusted odds ratio of 8.54 for patients who have a positive X-ray result as compared with a negative result (95% CI: 2.16 - 33.73).

Then, we set the log acid level to its mean value (4.2401) and a negative X-ray result, and obtained the following simulated results:

```
. setx mean
. setx xray 0
. simqi
```

Quantity of Interest	Mean	Std. Err.	[95% Conf.	Interval]
Pr (ni=abst)	.7428408	.0696632	.593598	.8614724
Pr (ni=pres)	.2571592	.0696632	.1385276	.406402

This means that a patient with a mean acid phosphatase level (69.415) and a negative X-ray result has a 26% probability of nodal involvement (95% CI: 14 – 41%).

For a patient with a mean acid phosphatase level that has a positive X-ray result we get:

```
. setx xray 1
. simqi
```

Quantity of Interest	Mean	Std. Err.	[95% Conf.	Interval]
Pr (ni=abst)	.2795374	.1074385	.1074838	.5101518
Pr (ni=pres)	.7204626	.1074385	.4898482	.8925162

In other words, this type of patient has a much higher probability, 72%, of nodal involvement (95% CI: 49 - 89%).

2. Survival of HIV+ patients. For another biomedical illustration, we take the data of an HMO-HIV+ survival study, described in (6) Table 1.1. After a confirmed diagnosis of HIV, patients were followed until death or until the subject was lost to follow-up. The main variable of interest, y, is survival time after confirmed diagnosis of HIV. A Kaplan-Meier survival curve, for drug users and non-users appears below.

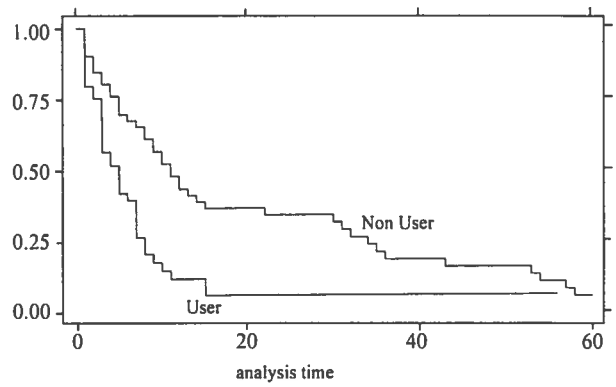

Figure 1. Kaplan-Meier survival estimates, by drug

In (6, chapter 8) three parametric models are estimated: *exponential*, *Weibull*, and *log-logistic* regression, with **x** = {patient's age, history of drug use} as covariates. All models made a good fit of the data; the authors observe that each model offers an accelerated time interpretation, and the estimates of effect are not different. Based on different model selection criteria, including Akaike's Information Criterion (*AIC*), the *Weibull* model (with the smallest *AIC*) seems to be the best fitting of the three. Therefore, we will use the *Weibull* regression model in our application of *Clarify*.

Following (6) we used the accelerated form of the Weibull model with shape parameter $\lambda = 1/\sigma$ and scale parameter $\gamma = \exp(-\beta_0/\sigma)$. Thus, the corresponding survival and hazard functions are:

$$S(t, \mathbf{x}, \beta, \sigma) \exp\{-t^{\lambda} \exp[-\mathbf{1}/\sigma) \cdot \mathbf{x}\,\beta]\}$$
$$h(t, \mathbf{x}, \beta, \lambda) = \lambda\gamma\,[t \exp(-\textstyle\sum \beta_j x_j)]^{\lambda-1} \cdot \exp(-\textstyle\sum \beta_j x_j).$$

The command *stsum* gives summary statistics of survival time: time at risk, incidence rate, number of subjects, and survival time percentiles, by drug (0 = Non-user, 1 = User).

```
. stsum, by(drug)
        failure _d: cens
  analysis time _t: time
```

		incidence	no. of	------- Survival time --------		
drug	time at risk	rate	subjects	25%	50%	75%
0	864	.0486111	51	5	11	34
1	272	.1397059	49	3	5	8
total	1136	.0704225	100	3	7	15

Clarify's *estsimp* command presents the results of the Weibull parametric model, with time as dependent variable, age and drug as covariates, and a censor indicator called *cens*.

```
. estsimp weibull time age drug, d(cens)
Weibull regression -- entry time 0
accelerated failure-time form                    Number of obs =    100
                                                 LR chi2(2)    =  52.05
Log likelihood = -128.50229                      Prob > chi2   = 0.0000
```

time	Coef.	Std.Err	z	P>|z|	[95% Conf.	Interval]
age	-.0908	.0136	-6.666	0.000	-.1174	- .0641
drug	-1.0492	.1890	-5.552	0.000	-1.4196	- .6788
_cons	6.1479	.5107	12.038	0.000	5.1469	7.1489
-ln(σ)	.1751	.0861	2.034	0.042	.0064	.3438
1/σ	1.1913	.1025			1.0064	1.4102
σ	.8394	.0722			.7091	.9936

In order to estimate the average survival time for a patient of median age (35 years) and a drug user:

```
. setx median
. setx drug 1
. simqi
```

we obtain the following results:

Quantity of Interest	Mean	Std. Err.	[95% Conf.	Interval]
E(time)	6.535744	.8927905	4.894694	8.355669

This means that a 35-year-old drug user, HIV positive patient is expected to survive between 4.9 and 8.4 months (with 95% probability).

For a 35 years old non-drug user, HIV positive patient we get the following:

```
. setx drug 0
. simqi
```

Quantity of Interest	Mean	Std. Err.	[95% Conf.	Interval]
E(time)	18.57398	2.34243	14.41033	23.82312

In other words, this type of patient is expected to live between 14.4 and 23.8 months (with 95% probability).

In terms of a difference of survival time, we get:

```
. simqi, fd(ev) changex(drug 0 1)
First Difference: drug 0 1
```

Quantity of Interest	Mean	Std. Err.	[95% Conf.	Interval]
dE(time)	-12.03823	2.524891	-17.37925	-7.469287

That is, a drug user is expected to survive between 7.5 and 17.4 fewer months than a non-user (with 95% probability).

Conclusion

We hope that the reader can appreciate the usefulness in biomedical research of *Clarify*, a simple but powerful technique developed by political science methodologists. *Clarify*, version 1.3, is capable to simulate a variety of models, including: regression, logit, probit, ordinal logit, ordinal probit, multinomial logit, Poisson, negative binomial, seemingly unrelated regression, and Weibull models. The use of statistical simulation has become a standard practice with the advent of fast and low cost modern computing. We have presented only part of what can be done with biomedical data using some simulation techniques, available via *Stata* and *Clarify*. From our standpoint, *Clarify* can be of much help on the statistical analysis of many biomedical problems. Other more powerful computer intensive methods have been developed, specially under the paradigm of *Monte Carlo - Markov Chain* for Bayesian and frequentist analysis which, also, are of much help on biomedical research [see, *e.g.* (7) and (8)].

It appears that recent exponents in biomedical research share the King and collaborators' approach, in favor of estimation as opposed to hypothesis testing. For example, (9, page 194) point out that:

The main thrust of the preceding sections has been to argue the inadequacy of statistical significance testing. The view that estimation is preferable to testing has been argued by many other statisticians and scientists ... Indeed, since statistical testing promotes so much misinterpretation, we recommend avoiding its use in epidemiologic presentations and research reports. Such avoidance requires that P-values (when used) be presented without reference to α-levels or "statistical significance."

We foresee that techniques like the one developed by King and his collaborators will close the gap between two scientific camps, biomedical researchers and political science methodologists, and that one will read the other literature more often.

Resumen

El propósito de esta monografía consiste en introducir el trabajo de un grupo de especialistas en ciencias políticas que crearon unos modelos para estimación y pronóstico estadístico. Se refiere al trabajo de Gary King y sus colaboradores en el Departamento de Gobierno de la Universidad de Harvard. Ellos desarrollaron un conjunto de técnicas de simulación computacional, junto a su programación de dominio público. Las fuentes básicas son su monografía, "*Making the most of statistical analyses: improving interpretation and presentation*," y su programa, *Clarify*. Con el advenimiento de computación rápida y de costo bajo, la utilización de

simulación estadística se ha convertido en una práctica común. El lector podrá apreciar la utilidad de *Clarify* en la investigación biomédica. Este programa puede ser de gran ayuda en el análisis estadístico de muchos problemas biomédicos.

Acknowledgements

We thank Messrs. Gary King, Michael Tomz, and Jason Wittenberg of Harvard University's Department of Government for the opportunity to work with their beta version of *Clarify*.

References

1. King G, Tomz M, Wittenberg J. Making the Most of Statistical Analyses: Improving Interpretation and Presentation. Am J Pol Sc 2000; 44: 341-355.
2. Tomz M, Wittenberg J, King G. Clarify: Software for Interpreting and Presenting Statistical Results. Version 1.2.1. Cambridge, MA: Harvard University. 2000 June 1. http://gking.harvard.edu/.
3. Brown BW. Prediction analyses for binary data. In: Miller RG, Efron B, Brown BW, and Moses LE, editors. Biostatistics Casebook. New York: John Wiley and Sons; 1980.
4. Collet D. Modelling Binary Data. London: Chapman & Hall; 1991.
5. Ryan TP. Modern Regression Methods. New York: John Wiley and Sons; 1997.
6. Hosmer DW, Lemeshow S. Applied Survival Analysis: Regression Modeling of Time to Event Data. New York: John Wiley and Sons; 1999.
7. Gilks WR, Richardson S, Spiegelhalter DJ, editors. Markov Chain Monte Carlo in Practice. New York: Chapman & Hall; 1996.
8. Rosen O, Jiang W, King G, Tanner MA. Bayesian and Frequentist Inference for Ecological Inference: the R x C case. Cambridge, MA: Harvard University; 2000 July 8; http://gking.harvard.edu/.
9. Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia, PA: Lippincott-Raven Publishers; 1998.