MACSS 30100: Problem Set 1

Tyler Amos

8 January, 2018

Part 1

a) Selected Article Synopsis: The article selected is from a 2016 issue of the *American Journal of Political Science*. It explores the issue of reporting bias and the impact of cell phone coverage on event reporting. It uses a Monte Carlo approach to develop certain parameters for the model (i.e., p), which it then applies to datasets on violent events and cell phone coverage in Afghanistan.

b) Citation: Weidmann, Nils B. "A Closer Look at Reporting Bias in Conflict Event Data". *American Journal of Political Science* 60, no. 1 (January 2016): 206-218. doi: 10.1111/ajps.12196

c) Model: $y = \beta_0 + \beta_1 x_1^p + log(k) + c + log(r) + n + \epsilon$

$y = \begin{cases} 1 & \text{if violence is reported in that location} \\ 0 & \text{if violence is not reported} \end{cases}$ $x_1 =$ cell phone coverage metric

$p = x \in [0, 1]$ $c = \frac{\text{number of coalition force casualities}}{k}$ $\epsilon =$ the error term

d) Table 1.1 Listing of Exogenous and Endogenous Variables

Variable	Rationale
Endogenous Variables	
Reported Violence (y)	This variable is the phenomenon of interest. It is hypothesized to be determined by other variables in the model.
Exogenous Variables	
Cell Phone Coverage (x_1)	The exogenous variable of interest, cell phone coverage. This variable is derived by the authors from the Mobile Coverage Explorer Dataset.
Likelihood of Reporting (p)	This variable allows the researcher to test different likelihoods of cell phone coverage affecting violent incident reporting.
Casualties (k)	This variable is thought to be a contributing factor to the likelihood of reporting an event. The more severe a violent incident (i.e., the more casualties) the more likely the event is to be reported.
Coalition Casualties (c)	This variable is calculated by the proportion of coalition force casualties to overall casualties. A higher proportion is hypothesized to render an event more newsworthy for international media and therefore more likely to be reported.
Remoteness (r)	This variable is used to control for periphery status of a location, which is hypothesized to be negatively correlated with reporting. It is used in three different ways by the author: i) the distance from the incident to the nearest settlement of 5,000 (town) or more; ii) the distance from the incident to the nearest settlement of 25,000 (city) or more; iii) population density at the event location.
Conflict in Neighboring Areas (n)	This variable adds information about conflict in neighboring areas, based on the assumption that a violent incident in one area may inspire violence nearby. It therefore accounts for spatial lag.

e) Table 1.2 Classification of Model

Classification	Rationale
Static	The model does not address changes in reporting over time.
Non-linear	The model makes use of *log* operators and logistic regression.
Stochastic	a) The model coefficients were estimated using logistic regression. b) The test and training data were random subsets of events drawn based on probabilities computed from the estimated coefficients. c) This simulation was repeated 1,000 times in a Monte Carlo simulation.

f) Missing Feature: I think it would be interesting to attempt this model with a continuous endogenous variable (e.g., $y \in [0, 1]$). The binary nature of the model was chosen by the authors to control for the effect of population density on violent incidents. As a result, an area is classified as either violent, or peaceful. However, the process of committing and reporting violent acts may in fact be usefully modelled as a differential equation, whereby an increase in violent acts in an area in the past can lead to more intense media and military focus on that area. Certain areas may gain a form of momentum in this way and be considered more newsworthy or strategically important. To do this, reports of violence would have to be taken as counts and then scaled to control for area populations. Then, a term could be added to the model to represent the rate of increase or decrease of reporting on violent events in a given location. This would also make this model dynamic rather than static.

2. a) Modelling an Individual's Decision to Wed: Let us assume that there are requirements for choosing to marry (e.g., interest, eligibility, partner's interest) and disqualifying factors (e.g., philosophical opposition to marriage, under the age of majority). Furthermore, let us focus on the decision to wed as driven by self-actualization/fulfillment, not economic or social considerations. Since the research question asks whether or not an individual will get married, we can assume the predictions should be valid for the individuals' entire lifetime from the present time. If we accept these assumptions, we can usefully model the decision to get married as a series of True/False statements. These statements can then be re-written as a deterministic model using boolean algebra.

$marriage = age \wedge eligibility \wedge interest_{self} \wedge (interest_{partner} \vee interest_{former}) \wedge (divorced \vee history) \wedge opposition$
$= (a)(e)(i_s)(\neg(\neg i_p \wedge \neg i_f))(\neg(\neg d \wedge \neg h))(o) \rightarrow \boldsymbol{m = (a)(e)(i_s)(1-(1-i_p)(1-i_f))(1-(1-d)(1-h))(o)}$

$marriage \begin{cases} 1 & \text{if get married} \\ 0 & \text{if not get married} \end{cases}$ $age = \begin{cases} 1 & \text{if the individual's age} \geq 18 \\ 0 & \text{if otherwise} \end{cases}$ $divorced = \begin{cases} 1 & \text{if previously married} \\ 0 & \text{if otherwise} \end{cases}$

$eligibility = \begin{cases} 1 & \text{if not currently married} \\ 0 & \text{if otherwise} \end{cases}$ $opposition = \begin{cases} 1 & \text{if no strong opposition to marriage} \\ 0 & \text{if otherwise} \end{cases}$

$interest_{self/former/partner} = \begin{cases} 1 & \text{if the individual, former partner, or current partner is interested in getting married} \\ 0 & \text{if otherwise} \end{cases}$

$history = \begin{cases} 1 & \text{if previously in, or desires to be in a long-term (} \geq 1 \text{ year), non-marriage relationship} \\ 0 & \text{if otherwise} \end{cases}$

b)Please note the dependent/endogenous variable m above.

c) Data could be simulated from this model because: Given all the parameters outlined above, the model could generate a prediction of whether or not the user would get married. Furthermore, the model could generate m-values with only some parameters. Specifically, only one of $\{i_p, i_f\}$ and one of $\{d, p\}$ would be required in order to make a prediction.

d & e, Part 1) Hypothesized key factors that influence the outcome of m: Age (a): This factor is a qualifying factor. While some individuals may marry under the age of 18, we excluded such situations (see assumptions, above).

Eligibility (e): The research question asks whether an individual **will** get married, thus we should exclude individuals who are already married. Only non-married individuals (single or in a non-married couple) are of interest in this model.

Interest: i_s) In order to marry, which we assume is for self-actualization, the individual must be interested in marriage. $i_p \vee i_f$) As marriage requires two individuals, we must include information on the current or former partners of the individual and their interest in marriage. This will indicate if the individual is in, or has been in relationships oriented towards long-term partnership.

Divorced Status/Relationship History: d) If the individual has been married before, this likely indicates an openess to marry again. h) If the individual has been, or would like to be in a long-term relationship, we can surmise they are also open to marriage.

Opposition (o): As hypothesized above, an individual who is opposed to marriage is unlikely to marry. To account for the individual's position evolving, this factor is conservatively applied - only when an individual strongly opposes marriage for moral/philosophical reasons would a 0 value be entered.

e) Part 2: Rationale for excluding other potential factors that influence the outcome of m: Factors which were considered, but ultimately not included in this model include: i) Wealth, which was excluded based on a heuristic evaluation - both wealthy and less-well-off individuals marry. ii) A continuous age factor was excluded for the purposes of keeping the model simple and time-invariant - the social importance of marriage has waned over time, for which the model would have to control. iii) A secular-religious dimension was excluded for concerns about generalizability to moderates. Highly religious individuals likely marry more than highly secular individuals, but I strongly doubt this would generalize to a large population. iv) No error term was included as this is designed to be a deterministic model.

f) Preliminary test for real-world factor significance: One way to test these factors for real-world significance would be to conduct a small-n survey through a platform such as Amazon MTurk. Participants would be asked questions to provide answers for the exogenous variables, as well as their current marital status. Some of the questions would have to ask married respondents to recall details from before they were married. This may provide some indication of the real-world factor significance. However, as this model is time-invariant, we must consider the possiblity that some individuals not married now will get married in the future.

A time-invariant way to test for factor significance would be to take a sample (as above) of a defined population for which we know, or can easily access the overall marriage rate (e.g., individuals living in Chicago). The researcher could use this dataset as a starting point for repeated (k) simulations with training and test sets derived from that sample. As we approach a large-k value, if the model approximates observed marriage rates acquired from census data, or a comparable source, we can confirm the factors have real-world significance. This approach might require some modification to the model in order to use stochastic techniques (e.g., Monte Carlo).