Sample Size Selection in Continuation-Ratio Logit Models

Andrew M. Raim[a*], **Thomas Mathew**[a,b], **Kimberly F. Sellers**[a,c], **Renee Ellis**[d], **Mikelyn Meyers**[d]

[a] Center for Statistical Research and Methodology, U.S. Census Bureau
[b] Dept of Mathematics and Statistics, University of Maryland, Baltimore County
[c] Dept of Mathematics and Statistics, Georgetown University
[d] Center for Behavioral Science Methods, U.S. Census Bureau

[*] Email: andrew.raim@census.gov

Summary

- Statistical agencies depend on data collected through contact with the public. Agencies may conduct experiments to study changes in response rates when contact procedures are altered.
- To study the effect of experimental factors on response rate, an experimenter might consider a logistic regression model with "success" taken to be a successful contact.
- When multiple contact attempts are made to the same respondent, effects may vary by attempt, and the experimenter may wish to account for this.
- Here we consider the continuation-ratio logit (CRL) model (e.g. Agresti, 2013), a sequential regression model where binary trials are carried out until either success or up to L failures.
- In this work, we consider the problem of sample size determination based on the Wald test of a general linear hypothesis.
- We present an illustration inspired by an experiment being considered for the 2020 Census Nonresponse Followup operation.

Continuation-Ratio Logit Model

- Let $W \in \{1, \ldots, L+1\}$ be a random variable with $\mathrm{P}(W = \ell) = p_\ell \prod_{b=1}^{\ell-1}(1 - p_b)$, given probabilities $\boldsymbol{p} = (p_1, \ldots, p_L)$.
- For $\ell \in \{1, \ldots, L+1\}$, we may then write

$$p_\ell = \frac{\mathrm{P}(W = \ell)}{\mathrm{P}(W = \ell) + \cdots + \mathrm{P}(W = L+1)} = \mathrm{P}(W = \ell \mid W \geq \ell)$$

  as a conditional probability of success during the ℓth trial, given that trials $1, \ldots, \ell - 1$ were unsuccessful (with $p_{L+1} \equiv 1$).
- We will write $W \sim \mathrm{CRL}_L(\boldsymbol{p})$ to describe this distribution. A CRL regression model for subjects $i = 1, \ldots, n$ is

$$W_i \sim \mathrm{CRL}_L(\boldsymbol{p}_i), \quad \mathrm{logit}(p_{i\ell}) = \boldsymbol{x}_{i\ell}^\top \boldsymbol{\beta}, \quad \ell = 1, \ldots, L.$$

- The likelihood, score, and information matrix are

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{\ell=1}^{L+1} \left[p_{i\ell} \prod_{b=1}^{\ell-1} (1 - p_{ib}) \right]^{I(w_i = \ell)},$$

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{\ell=1}^{L+1} \Big[I(w_i = \ell) - I(w_i \geq \ell) G(\eta_{i\ell}) \Big] \boldsymbol{x}_{i\ell},$$

$$\mathcal{I}(\boldsymbol{\beta}) = \boldsymbol{X}^\top \boldsymbol{D}_{\boldsymbol{\beta}} \boldsymbol{X}, \quad \boldsymbol{D}_{\boldsymbol{\beta}} = \mathrm{Diag} \left\{ g(\boldsymbol{x}_{i\ell}^\top \boldsymbol{\beta}) \prod_{b=1}^{\ell-1} [1 - p_{ib}] \right\},$$

  where $G(x)$ is the inverse logit function and $g(x) = G'(x)$.

Recoding

- An observed w_i can be recoded using L binary variables $(y_{i1}, \ldots, y_{iL})$, with $y_{i\ell} = 1$ if $\ell = w_i$, $y_{i\ell} = 0$ if $\ell < w_i$, and $y_{i\ell} = \mathrm{NA}$ if $\ell > w_i$.
- The CRL likelihood is equivalent to logistic regression based on observations $\{y_{i\ell}\}$ with NA values dropped.
- Logistic regression can be used to compute the CRL MLE $\hat{\boldsymbol{\beta}}$, but associated variance estimates will generally differ from $\mathcal{I}^{-1}(\hat{\boldsymbol{\beta}})$.

Testing Problem

- Given a known matrix $\boldsymbol{C} \in \mathbb{R}^{q \times d}$ with rank $q \leq d$ and known vector $\boldsymbol{c}_0 \in \mathbb{R}^q$, consider the general linear hypotheses

$$H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{c}_0 \quad \text{vs.} \quad H_1 : \boldsymbol{C}\boldsymbol{\beta} \neq \boldsymbol{c}_0.$$

- A Wald test with significance level α is

  Reject H_0 if $\mathcal{T} > \chi^2_q(1 - \alpha)$, where
  $\mathcal{T} = (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{c}_0)^\top (\boldsymbol{C}\mathcal{I}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{C}^\top)^{-1} (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{c}_0)$
  and $\chi^2_q(\gamma)$ is the γ quantile of χ^2 with q degrees of freedom.
- For large samples, the power of the test is approximately

$$\varpi = 1 - F_{\mathcal{T}}(\chi^2_q(1 - \alpha); q, \psi(\boldsymbol{\beta})), \quad \text{with}$$
$$\psi(\boldsymbol{\beta}) = (\boldsymbol{C}\boldsymbol{\beta} - \boldsymbol{c}_0)^\top (\boldsymbol{C}\mathcal{I}^{-1}(\boldsymbol{\beta})\boldsymbol{C}^\top)^{-1} (\boldsymbol{C}\boldsymbol{\beta} - \boldsymbol{c}_0),$$

  where $F_{\mathcal{T}}(w; q, \psi)$ is the CDF of χ^2 with q degrees of freedom and non-centrality parameter ψ.
- **Question:** When planning an experiment where the objective is to carry out this test, how to select an adequate sample size?

Computing Power

- For this work, we assume the "nuisance" parameter is known. That is, there is a $\boldsymbol{B} \in \mathbb{R}^{(q-d) \times d}$ so that the matrix $(\boldsymbol{B}^\top \; \boldsymbol{C}^\top)^\top$ is non-singular and $\boldsymbol{B}\boldsymbol{\beta} = \boldsymbol{b}_0$ is known.
- Covariates $\mathcal{X} = \{\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{nL}\}$ are assumed to be known.
- We characterize the departure from H_0 using $\Delta \geq 0$, and let $\mathcal{S}(\boldsymbol{c}_0, \boldsymbol{b}_0, \Delta) = \{\boldsymbol{\beta} \in \mathbb{R}^d : \|\boldsymbol{C}\boldsymbol{\beta} - \boldsymbol{c}_0\| = \Delta, \boldsymbol{B}\boldsymbol{\beta} = \boldsymbol{b}_0\}$.
- For a given Δ, the power $\varpi(\boldsymbol{\beta})$ may vary for $\boldsymbol{\beta} \in \mathcal{S}(\boldsymbol{c}_0, \boldsymbol{b}_0, \Delta)$.
- To be conservative, we take the power to be $\varpi(\tilde{\boldsymbol{\beta}})$ using

$$\tilde{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathcal{S}(\boldsymbol{c}_0, \boldsymbol{b}_0, \Delta)} \psi(\boldsymbol{\beta}).$$

- Using an appropriate transformation, this becomes an unconstrained minimization problem.

Determining Sample Size

Given covariate data $\mathcal{X}$, an investigation to determine sample size can be carried out as follows:

1. Determine samples $\mathcal{J}_1, \ldots, \mathcal{J}_m \subseteq \{1, \ldots, n\}$ of increasing size which are viable for the experiment.
2. Determine a grid $\{\Delta_1, \ldots, \Delta_r\}$ of effect sizes to consider.
3. For each combination of $\Delta \in \{\Delta_1, \ldots, \Delta_r\}$ and $\mathcal{J} \in \{\mathcal{J}_1, \ldots, \mathcal{J}_m\}$, compute $\tilde{\boldsymbol{\beta}}$ to find power $\varpi(\tilde{\boldsymbol{\beta}})$.

We can then select the smallest $j \in \{1, \ldots, m\}$ which achieves sufficient power and sensitivity.

Sparse Categories

- Care must be taken when selecting the number of attempts L to model; too many can lead to an issue of sparse categories.
- To see this, consider the simple model $W \overset{\text{ind}}{\sim} \mathrm{CRL}_5(p, \ldots, p)$; some probabilities $\mathrm{P}(W = \ell) = p(1 - p)^{\ell - 1}$ are shown below.

	Attempt					
p	1	2	3	4	5	6+
0.05	0.05	0.048	0.045	0.0429	4.073E-2	7.738E-1
0.10	0.10	0.090	0.081	0.0729	6.561E-2	5.905E-1
0.25	0.25	0.188	0.141	0.1055	7.910E-2	2.373E-1
0.40	0.40	0.240	0.144	0.0864	5.184E-2	7.876E-2
0.60	0.60	0.240	0.096	0.0384	1.536E-2	1.024E-2
0.75	0.75	0.188	0.047	0.0117	2.930E-3	9.766E-4
0.90	0.90	0.090	0.009	0.0009	9.000E-5	1.000E-5
0.95	0.95	0.048	0.002	0.0001	5.938E-6	3.125E-7

- Large p may give very small $\mathrm{P}(W = \ell)$ for ℓ later in the sequence.

References

Alan Agresti. *Categorical Data Analysis*. Wiley, 3rd edition, 2013.

R. Ellis, P. Goerman, K. Kephart, A. C. Fobia, A. S. Giron, M. Meyers, R. Terry, L. Fernandez, F. Lineback, M. Berger, A. Bruce, and E. Jensen. Research on Coverage of Underrepresented Populations in Anticipation of a Records-Based Census, 2018. 2020 Census: Evaluation, Experiment, and Research and Testing Study.

A. M. Raim, T. Mathew, K. F. Sellers, R. Ellis, and M. Meyers. Experiments on Nonresponse Using Sequential Regression Models, 2020+.

U.S. Census Bureau. 2020 Census Detailed Operational Plan for: 18. Nonresponse Followup Operation (NRFU), July 2019. https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/planning-docs/NRFU-detailed-op-plan.html.

Sample Size Selection in Continuation-Ratio Logit Models

Andrew M. Raim[a*], **Thomas Mathew**[a,b], **Kimberly F. Sellers**[a,c], **Renee Ellis**[d], **Mikelyn Meyers**[d]

[a] Center for Statistical Research and Methodology, U.S. Census Bureau
[b] Dept of Mathematics and Statistics, University of Maryland, Baltimore County
[c] Dept of Mathematics and Statistics, Georgetown University
[d] Center for Behavioral Science Methods, U.S. Census Bureau

[*] Email: andrew.raim@census.gov

Illustration

- The following illustration is based on an actual experiment being considered for 2020 Census Nonresponse Followup (NRFU).
- A training module has been developed to guide enumerators in administering the census questionnaire to Spanish-speaking households. The objective is to improve consistency in messaging and in the usage of official translations (Ellis et al., 2018).
- **Main study question:** does the new training significantly affect response rates for Spanish-speaking households? To answer this question, we describe a statistical experiment under CRL and investigate sample size.
- Experimental subjects are Spanish-speaking households in NRFU, not known with certainty before the operation.
- Control ("no training") or experimental ("training") treatments are assigned to enumerators at the level of Area Census Office (ACO). For this discussion, an ACO can be considered to be an administrative grouping of nearby tracts.
- In the 2020 Census, cases will be assigned dynamically based on enumerator availability and workloads (U.S. Census Bureau, 2019). A household may be visited by multiple enumerators.
- Fourteen ACOs were pre-selected from several metropolitan statistical areas (MSAs) in Dallas, Houston, and Los Angeles. Displayed data are from 2019 Planning Database.

		Percent		HH Counts	
Area	Group	Spanish	Selfresp	Total	Target
Dallas	Ctrl	6.8	62.8	352,347	11,900
Dallas	Ctrl	14.2	48.5	293,170	24,847
Dallas	Ctrl	10.4	57.4	337,574	19,828
Dallas	Ctrl	24.9	41.2	277,452	43,271
Dallas	Ctrl	11.6	55.6	335,557	23,521
Dallas	Ctrl	4.0	66.3	482,153	8,084
LA	Ctrl	13.9	49.5	441,726	35,989
Houston	Expt	21.0	44.1	253,932	33,305
Houston	Expt	10.1	47.9	278,782	18,412
Houston	Expt	15.8	44.0	282,424	31,434
Houston	Expt	21.6	41.3	240,950	36,575
Houston	Expt	20.0	40.7	238,144	32,587
Houston	Expt	8.0	61.3	268,572	9,525
LA	Expt	16.1	48.5	496,564	50,740
Total				4,579,347	380,018

- **Main sample size question:** is this initial selection of ACOs adequate for the experiment?

Illustration II

- Control and experimental ACOs have been geographically separated to avoid "contamination" in the study, where households are visited by both types.
- Let the number of contacts needed for a response be

$$W_{ijk} \sim \text{CRL}_L(\boldsymbol{p}_{ijk}), \quad \begin{aligned} i &= i, \ldots, I = 7, \\ j &= 1, \ldots, J = 2, \\ k &= 1, \ldots, K_{ij}, \end{aligned}$$

  for the kth Spanish-speaking NRFU household within the ith ACO which received the jth treatment.
- The control and experimental treatments are indexed by $j = 1$ and $j = 2$, respectively.
- A rough estimate of K_{ij} is obtained from the 2019 Planning Database using

$$\text{HH\_Target} = \text{HH\_Total} \times \frac{\text{Pct\_Spanish}}{100} \times \frac{1 - \text{Pct\_Selfresp}}{100}.$$

- For probabilities of a response at each attempt, we consider

$$\begin{aligned} \text{logit}(p_{ijk\ell}) &= \mu + \tau_j + \delta_\ell + (\tau\delta)_{j\ell} \\ &= \boldsymbol{s}_{j\ell}^\top \boldsymbol{\beta}, \end{aligned}$$

  assuming a parameterization

$$\boldsymbol{\beta} = \left(\mu, \tau_1, \delta_1, \ldots, \delta_{L-1}, (\tau\delta)_{11}, \ldots, (\tau\delta)_{1,L-1}\right) \in \mathbb{R}^{2L},$$

  assuming the constraints

$$\sum_{j=1}^{J} \tau_j = 0, \quad \sum_{\ell=1}^{L} \delta_\ell = 0, \quad \sum_{j=1}^{J} (\tau\delta)_{j\ell} = 0, \quad \sum_{\ell=1}^{L} (\tau\delta)_{j\ell} = 0.$$

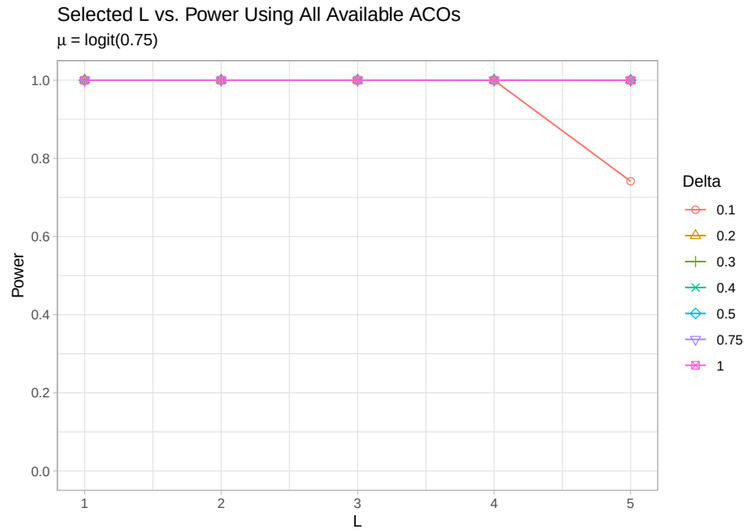
- The effects are: an intercept term μ, effects due to the treatment τ_j, effects due to contact attempt δ_ℓ, and effects $(\tau\delta)_{j\ell}$ due to treatment-attempt interaction.
- Sample size will be based on a significance level $\alpha = 0.10$ test for the presence of any treatment effects

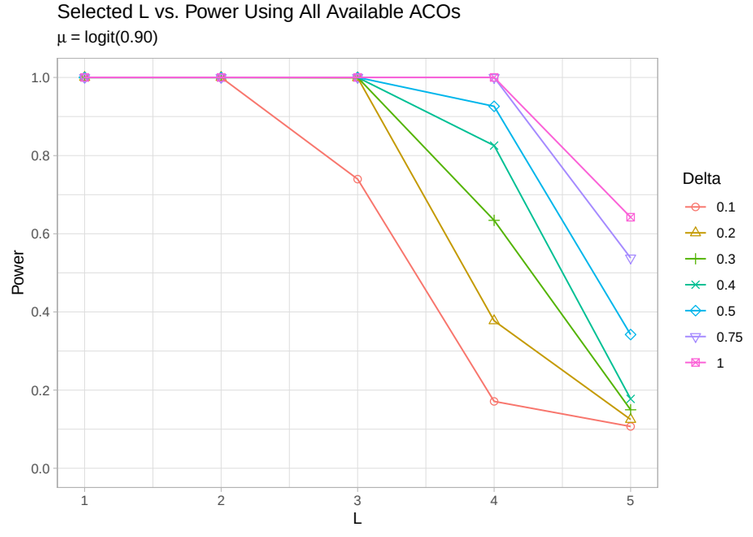
$$\begin{aligned} &H_0 : \tau_1 = 0, \delta_\ell, (\tau\delta)_{1\ell} = 0, \quad \text{for } \ell = 1, \ldots, L-1 \\ &\text{vs. } H_1 : \text{Not}, \end{aligned}$$

  so that μ is the nuisance parameter.
- Rewriting this as a general linear hypothesis, we may investigate the relationship between the sample size, the effect size Δ, the power ϖ, and μ.

Results

- Simulation results confirm that the non-central χ^2 power approximation breaks down when categories become sparse. Furthermore, MLE computation becomes more likely to fail.
- Assuming the non-central χ^2 power approximation is appropriate, power declines sharply with increasing L when the baseline response effect μ is large; i.e., as $\text{logit}^{-1}(\mu)$ approaches 1.





- Assuming $\mu \leq \text{logit}(0.90)$, we would suggest $L = 3$. This gives power $\varpi \approx 0.77$ to detect effect size $\Delta = 0.1$ with the 14 ACOs.
- Details will be provided in Raim et al. (2020+, in preparation).
- Mixed effects and handling of unknown quantities—including K_{ij} and nuisance parameters—will be considered in future work.