Investigating the Effect of Design Weights in a Complex Survey Design

Jonathan and Dong , University of the Fraser Valley
abelivea@sfu.ca, zhengs@sfu.ca

Introduction

Objective: Determine the risk factors for hypertension among Canadians with and without design weights. Investigate whether these risk factors depend on gender or age.

Study: Statistics Canada conducted Cycle 3 of their Canadian Health Measures Survey from 2012-2013.

- A Mobile Examination Centre (MEC) was used to take direct physical measurements from approximately 3000 Canadians across the ten provinces.
- Survey weights were assigned to each study participant to insure that the sample represented the target population.
- Due to the complexity of the stratified three-stage sample design, bootstrap weights were created to estimate the variance of estimators.

Data: The following measurements were taken directly from study participants, and these were considered to be potential risk factors for hypertension in our analysis:

- CLINIC ID: Unique record identifiers.
- SMK 12: Current smoking status: 1 daily; 2 occasional; 3 non-smoker.
- CLC SEX: Sex at clinic visit: 1 male, 2 female.
- CLC AGE: Age in years at clinic visit: 20 to 79.
- HWMDBMI: Body mass index in kg/m2. Based on measured height and weight.
- HIGHBP: Categorized hypertensive: 1 yes, 2 no. A respondent is categorized as hypertensive if he/she has SPB $\geq$ 140 mmHg or DBP $\geq$ 90 mmHg or is treated for hypertension (taking medication and/or been diagnosed as hypertensive by a medical professional in the past 6 months).
- LAB BCD: Blood cadmium in nmol/L.
- LAB BHG: Blood mercury in nmol/L.

Notation

Data: Let i denote the study participant, and j the explanatory variable. The data can be summarized into the following quantities:

- n: Sample size
- y_i: Vector whose ith element is the HIGHBP value of the ith study participant
- x_i: Vector whose jth element is the observed value of the jth covariate for the ith study participant
- w_i: Vector whose ith element is the survey weight for the ith study participant
- $g()$: Link function
- μ_i: Vector whose ith element equals the expected value of HIGHBP for the ith study participant
- $\hat{\beta}_j$: Vector whose jth element is the estimated coefficient of the jth explanatory variable

Cleaning the Data

Censored Data: The variables LAB BCD and LAB BHG both contained data that was below their respective Limit of Detection (LOD).

- This censored data was imputed with values randomly drawn from the interval [0, LOD].

Model

Unweighted Model: A generalized linear model without using design weights was first created.

- The response HIGHBP was modelled as independent categorical r.v.'s with parameters u_i.
- The logit link function was used to model the relationship between u_i and the explanatory variables.

The parameter coefficients β are estimated with their MLE's, which satisfy the following equation:

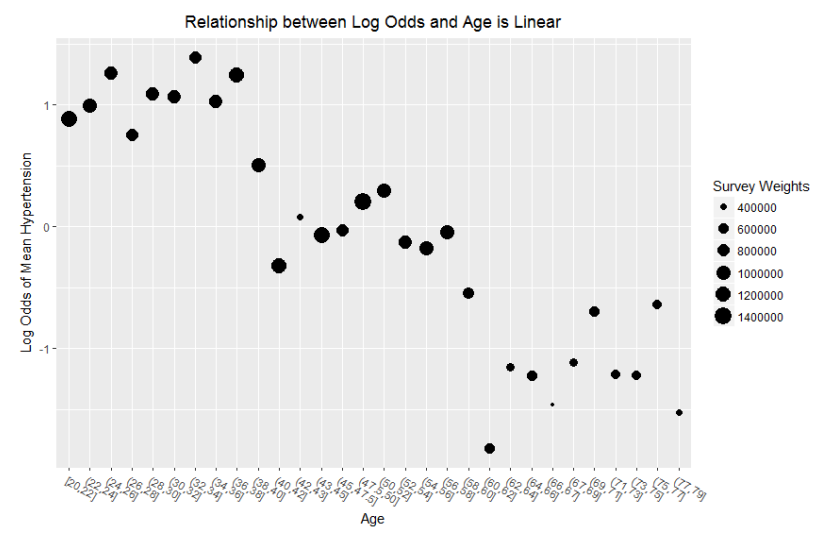
$$U(\hat{\beta}) = \sum_{i=1} nx_i \frac{1}{g'(\mu_i)V(\mu_i)}(y_i - \mu_i(\hat{\beta})) = 0 \qquad (1)$$

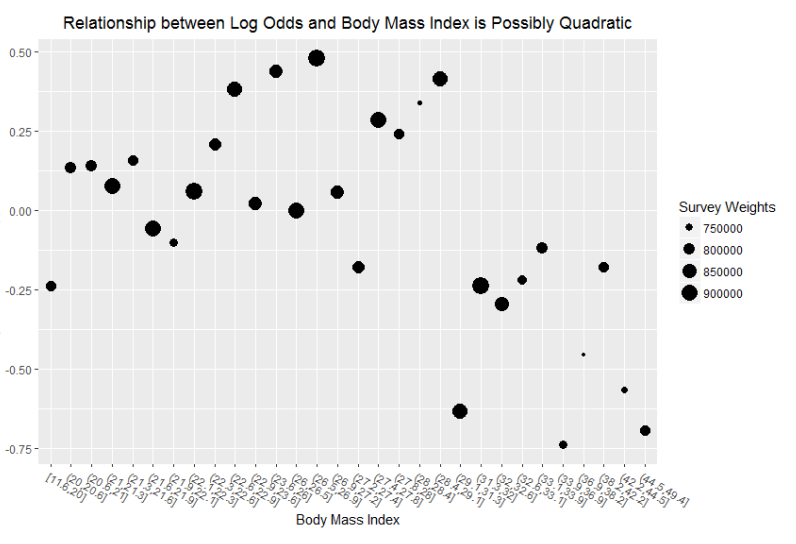
Weighted Model: A model using the design weights was then created. Following Lumley..., the score function above was altered to incorporate design weights, and this was solved to get the parameter estimates:

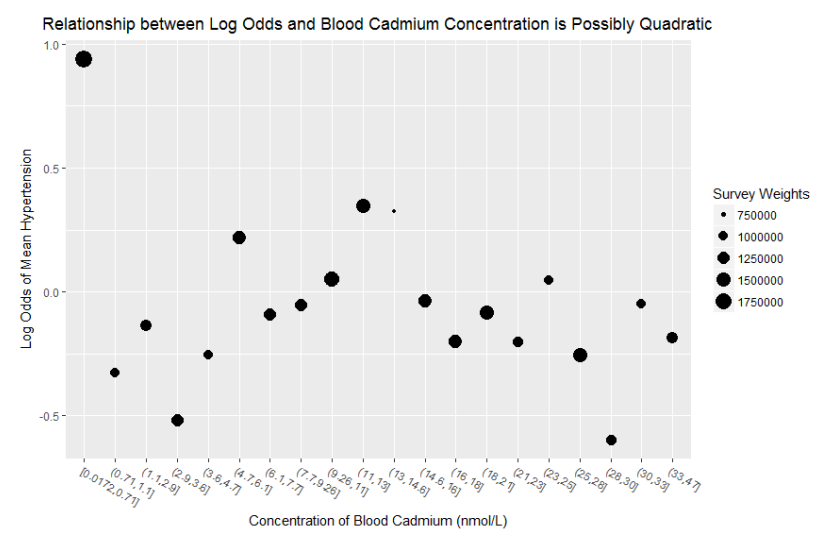
$$U(\hat{\beta}) = \sum_{i=1} nw_i x_i \frac{1}{g'(\mu_i)V(\mu_i)}(y_i - \mu_i(\hat{\beta})) = 0 \qquad (2)$$

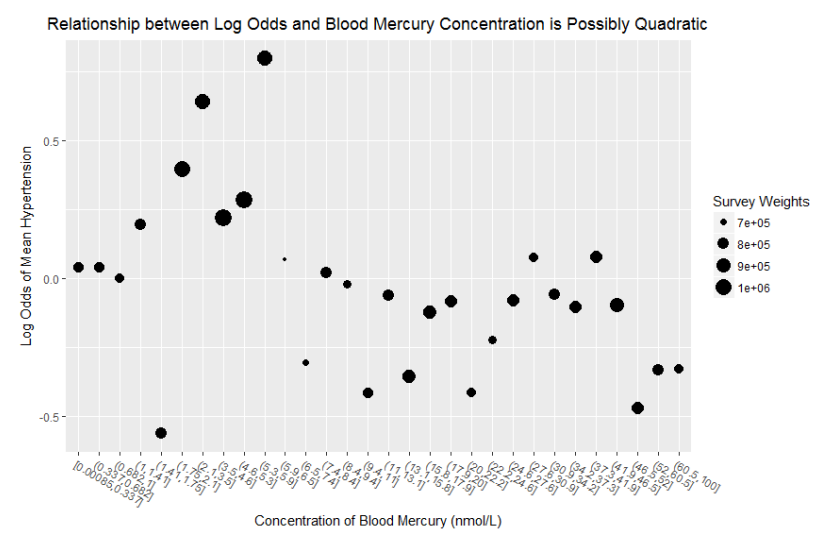
Exploratory Analysis

Purpose: To determine whether higher-order terms should be included in the models. In these plots, the continuous variables were made into categorical variables to avoid observed hypertension proportions of 0 or 1.









These graphs suggest that including second-order terms for HWMDBMI, LAB BCD, and LAB BHG in the model could be informative, but isn't necessary for CLC AGE.

Hypothesis Testing Theory

Null Hypothesis: Suppose β is partitioned by $(\beta_{(1)}, \beta_{(2)})$. The null hypothesis is H_0: $\beta_{(1)} = 0$, where the dimension of $\beta_{(1)}$ is q.
Unweighted Model: For the unweighted model, the log-likelihood ratio statistic was used to test H_0. Under the null hypothesis, this statistic follows a Chi-Square distribution with q degree of freedom.
Weighted Model: A different method was now needed because the likelihood equations do not exist for the weighted model. Following Lumley..., the standard likelihood function was altered to include weights:

$$\ell(\hat{\beta}) = \sum_{i=1} nw_i \log f(y_i|x_i;\beta) \qquad (3)$$

where f is the pdf of HIGHBP. The working likelihood ratio test statistic was then defined by Lumley as

$$\tau = 2\left[\ell(\hat{\beta}) - \ell(\hat{\beta}_0)\right] \qquad (4)$$

where $\hat{\beta}_0$ is the solution to the weighted score function when $\beta_{(1)}$ = 0. Finally, τ has a rather complex distribution under H_0, but the saddlepoint approximation (Kuonen, 1999) can be used to approximate it.

Methods (cont'd)

Interpretation: Element (s,t) of the matrix $E(R_{ijDF})$ is the expected number of double tagged fish released in year i in region s $(\tilde{N}_{isDD})$ that survived $(\tilde{\phi}_i)$ and migrated (Q_i) and were not captured $(1-\tilde{p}_i)$ during year i. Further, they survived, migrated and were not captured up to the end of year $j-1$. Then, during year j, they survived, migrated to region t and finally were captured (p_j) with only a front tag (Λ_{ijDF}) and were reported (λ_{jDF}).

Tag Loss Model: Following Brattey and Cadigan (2006) [?], we model Λ_{ilF} and Λ_{ilG} using Kirkwood's parametric model. That is, for $h = F$ or B,

$$\Phi_{ilh} = \left[\frac{\beta_{1h}}{\beta_{1h} + \beta_{2h}(l-i)}\right]^{\beta_{2h}}, \beta_{1h} > 0, \beta_{2h} > 0. \qquad (5)$$

Constraining parameters: Parameters that represent probabilities are constrained between 0 and 1 using a logit transformation. Kirkwood's parameters in () are constrained > 0 using an exponential transformation.

Point & Variance Estimation: Maximum likelihood and Delta method

Model Selection: Due to the complexity of the model and the sparse, we have been able to fit only very few models with success. Therefore, we will only present the most sophisticated model we have been able to fit. We would suggest using QAIC to compare different models.

Computing: We use R version ... Our code is general for any time period length or number of regions. We use design matrices allowing to constraint some parameters to be equal. (Newton-Raphson algorithm)

Results

Exploitation rates: Table showing estimate(SE) in % per region and year. (*) : Unable to recover SE numerically

	East	Off	South	West
1997	0.6(0.2)	0.9(0.8)	5.5(*)	8.1(12.5)
1998	5.5(0.9)	2.3(0.5)	5.9(*)	2.3(0.7)
1999	11.3(1.6)	3.3(0.6)	13.4(*)	2.7(0.7)
2000	5.8(0.5)	4.2(*)	12.5(0.6)	0.9(0.2)
2001	7.8(0.6)	3.2(*)	11.2(0.6)	1.7(0.3)
2002	7.6(0.6)	2.6(*)	8.9(0.5)	0.9(0.2)
2003	14.8(1.7)	2.7(0.5)	8.2(0.1)	0.3(0.1)
2004	3.8(0.6)	3.2(0.6)	7.4(0.2)	0.4(0.2)
2005	3.1(0.5)	6.4(1.2)	7.4(0.2)	0.5(0.2)
2006	5.9(0.5)	5.4(0.7)	8.1(0.8)	0.3(0.2)
2007	3.1(0.3)	21.4(1.4)	4.1(0.5)	1.1(0.4)
2008	3.2(0.3)	10.3(1.2)	7.1(0.8)	1.1(0.4)
2009	2.6(0.3)	7.8(*)	5.0(0.4)	1.0(0.4)
2010	1.7(0.2)	6.9(*)	5.6(0.7)	0.7(0.3)
2011	1.0(0.2)	2.8(0.6)	2.6(0.4)	0.4(0.3)

Tag Retention:
Estimated Cumulative Tag Retention Probabilities (Kirkwood's Model)

kirkwood.png

Migration:

	E	O	S	W
E	94.5	0.6	4.7	0.1
O	0.6	47.6	5.2	46.7
S	3.0	3.3	93.2	0.6
W	0.3	1.9	0.2	97.5

Reporting rates: For fish recovered as S, DF or DB, it varies from 18 % to 99 % with SE's ranging from 3 % to 18 %. For fish recovered as DD, it varies from 7 % to 100 % but many SE's are large and probably indicate identifiability problem. See "further work".

Results (cont'd)

Annual Survival Rate : 74.6 %, SE=0.4. We investigated the effect of doubling the natural death rate on exploitation rates by fixing the survival probability to 49.3 % in our modelï£¡ Following table shows estimate(SE) in % per region and year. (*) : Unable to recover SE numerically

	East	Off	South	West
1997	0.6(0.2)	0.9(0.8)	5.5(*)	8.1(12.5)
1998	5.5(0.9)	2.3(0.5)	5.9(*)	2.3(0.7)
1999	11.3(1.6)	3.3(0.6)	13.4(*)	2.7(0.7)
2000	5.8(0.5)	4.2(*)	12.5(0.6)	0.9(0.2)
2001	7.8(0.6)	3.2(*)	11.2(0.6)	1.7(0.3)
2002	7.6(0.6)	2.6(*)	8.9(0.5)	0.9(0.2)
2003	14.8(1.7)	2.7(0.5)	8.2(0.1)	0.3(0.1)
2004	3.8(0.6)	3.2(0.6)	7.4(0.2)	0.4(0.2)
2005	3.1(0.5)	6.4(1.2)	7.4(0.2)	0.5(0.2)
2006	5.9(0.5)	5.4(0.7)	8.1(0.8)	0.3(0.2)
2007	3.1(0.3)	21.4(1.4)	4.1(0.5)	1.1(0.4)
2008	3.2(0.3)	10.3(1.2)	7.1(0.8)	1.1(0.4)
2009	2.6(0.3)	7.8(*)	5.0(0.4)	1.0(0.4)
2010	1.7(0.2)	6.9(*)	5.6(0.7)	0.7(0.3)
2011	1.0(0.2)	2.8(0.6)	2.6(0.4)	0.4(0.3)

Further Work

Length: To be incorporated in the analysis because exploitation rate is known to vary by length. This involves growth curve estimation, see method in Cadigan and Brattey (2001) [?] which was applied to a subset of the 1997-2000 data.

Population structure: Use a latent-state model (eg. bayesian approach) to distinguish between resident inshore and migrant offshore cods.

Reporting rates: As in [?], estimate the odds ratio of reporting double vs single low-reward tags, rather than estimating λ_{DD} (results in improved precision in estimating double tag reporting rates and reduced number of parameters).

Model Sophistication: Fit models successfully using smaller time scale (season, month) and areas.

2-step MLE: Preliminary developements suggest that likelihood can be broken in 2 pieces that can be maximized successively. This allows to estimate tag retention and reporting rates separately, reducing the complexity of the problem.

References

[1] Brattey, J. and Cadigan N.G. (2006). Reporting and Shedding Rate Estimates From Tag-Recovery Experiments on Atlantic Cod (Gadus Morhua) in Coastal Newfoundland. *Can. J. Fish. Aquatic Sc.*, 63(9), 1944-1944.

[2] Cadigan N.G. and Brattey, J. (2001). A Nonparametric Von Bertalanffy Model for Estimating Growth Curves of Atlantic Cod. ICES CM 2001/O:18. Available from: http://www.ices.dk/products/CMdocs/2001/O/O1801.pdf. Accessed May 2012.

[3] Cowen, L. et al. (2009). Estimating Exploitation Rates of a Migrating Population of Yellowtail Flounders Using Multi-State Mark-Recapture Methods Incorporating Tag-Loss and Variable Reporting Rates. *Can. J. Fish. Aquatic Sc.*, 66, 1245-1273.

Acknowledgements

We are indebted and very grateful to Laura Cowen for sharing her code from [?]. Special thanks to our mentor, Carl Schwarz. The authors acknowledge the support from NSERC.