Investigating the Effect of Design Weights in a Complex Survey Design

Jonathan Fitz and Dong Liang, University of the Fraser Valley
Jonathan.Fitz@student.ufv.ca, Dong.Liang@student.ufv.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 survey that utilizes two different kinds of weights.

- 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.

Response: HIGHBP: Categorized hypertensive: 1 yes, 2 no.

Potential Risk Factors: The following measurements were taken directly from study participants using the MEC:

- SMK: Smoking status: 1 daily; 2 occasional; 3 non-smoker.
- SEX: Sex at clinic visit: 1 male, 2 female.
- AGE: Age in years at clinic visit: 20 to 79.
- BMI: Body mass index in kg/m2.
- BCD: Blood cadmium in nmol/L.
- BHG: Blood mercury in nmol/L.

Notation

Data: Let i denote the study participant, and j the explanatory variable. Other useful terms are defined below:

- y: Vector whose ith element is the HIGHBP value of the ith study participant.
- X: Design matrix whose ijth element is the observed value of the jth covariate for the ith study participant.
- w: Vector whose ith element is the survey or bootstrap weight for the ith study participant .
- n, V: Sample size and variance respectively.
- $f(X)$: The probability mass function for HIGHBP.
- μ: Vector whose ith element equals the expected value of HIGHBP for the ith study participant.
- β: Vector whose jth element is the coefficient of the jth explanatory variable.
- $g(\mu)$: Link function to model the relationship between μ and the explanatory variables.
- q: Number of parameters set to zero in null hypothesis.

Cleaning the Data

Censored Data: The variables BCD and 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].
- Popular alternatives include imputing using a single value such as LOD or LOD/2, but our approach should better account for the variance of the true (unknown) data.

Cleaning the Data Continued

Missing Data: Roughly 6% of observations across four variables contained missing data.

- This missing data was imputed using K-nearest neighbors (KNN), and 5-fold cross-validation was used to pick K.
- Four KNN models were built, each having as the response one of the four variables with missing data.
- The training set for these models was the set of all observations that did not contain any missing data.

Models

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

- HIGHBP was modelled as binary response r.v.'s with parameters μ, and the logit was chosen for the link function.

The model coefficients β were estimated using maximum likelihood, which satisfy the following equation:

$$U(\hat{\beta}) = \sum_{i=1}^{n} x_i \frac{1}{g'(\mu_i)V(\mu_i)}\left(y_i - \mu_i(\hat{\beta})\right) = 0 \tag{1}$$

Weighted Model Estimators: Following Lumley and Scott (2017), the score function above was altered to incorporate design weights:

$$U(\hat{\beta}) = \sum_{i=1}^{n} w_i x_i \frac{1}{g'(\mu_i)V(\mu_i)}\left(y_i - \mu_i(\hat{\beta})\right) = 0 \tag{2}$$

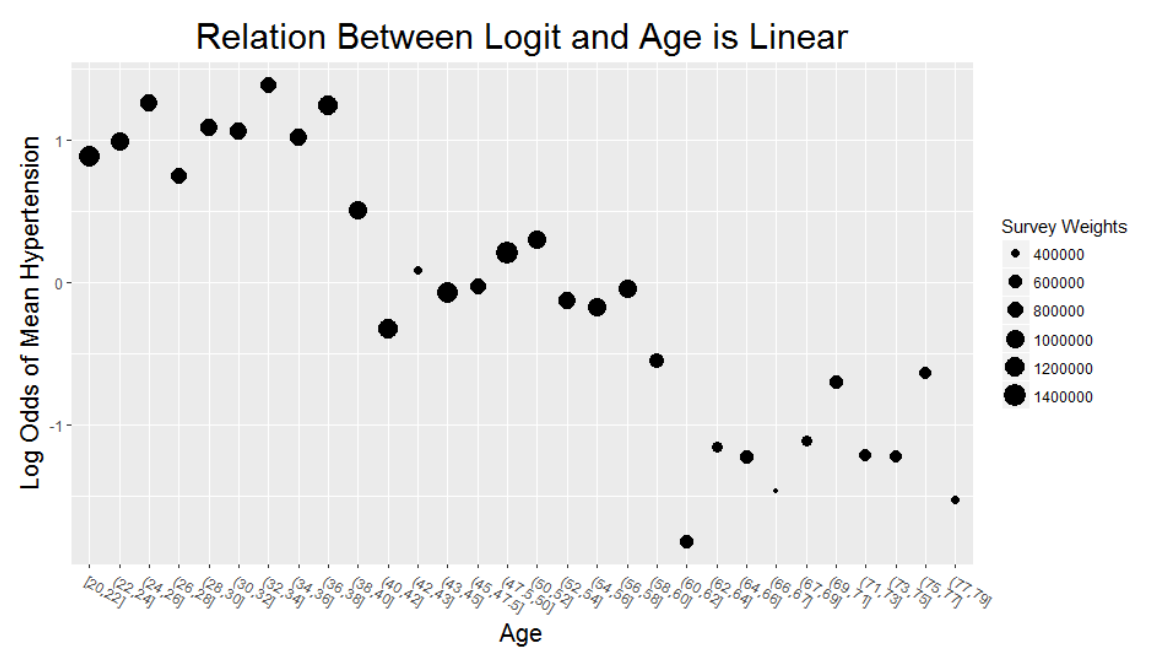
Estimating Variance:

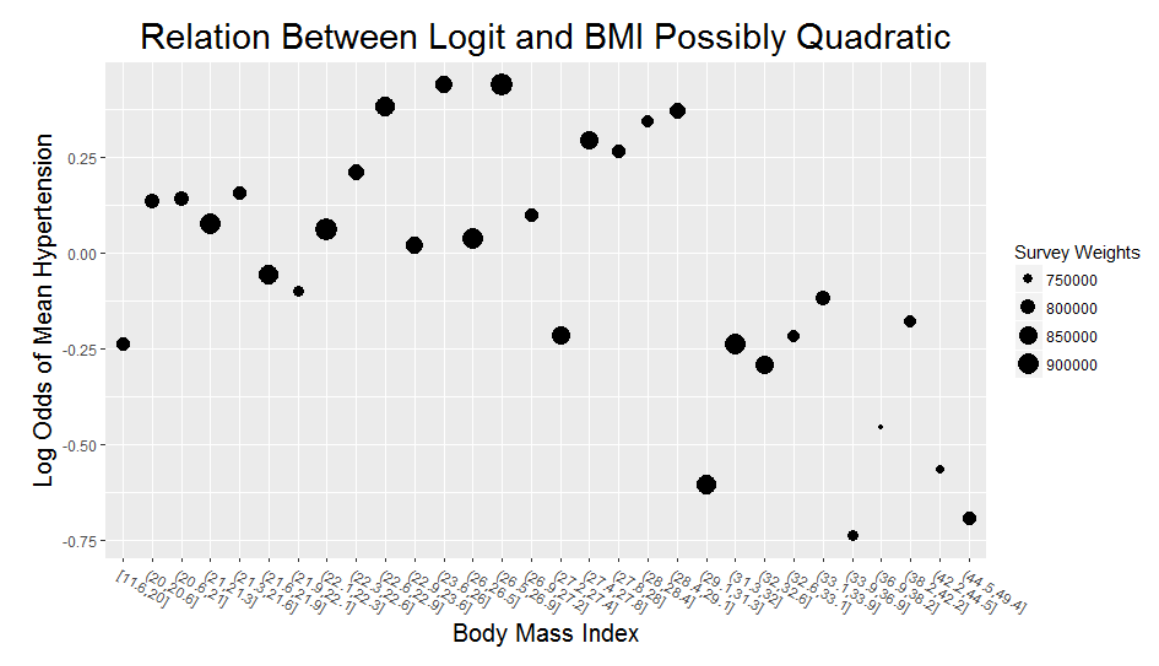
- For unweighted model, the variance of model coefficients were estimated using the inverse of the Information matrix.
- For weighted model, each of the 500 bootstrap weights was used with (2) to calculate 500 more estimates. The variance of these estimates was the estimated variance of the model coefficients.

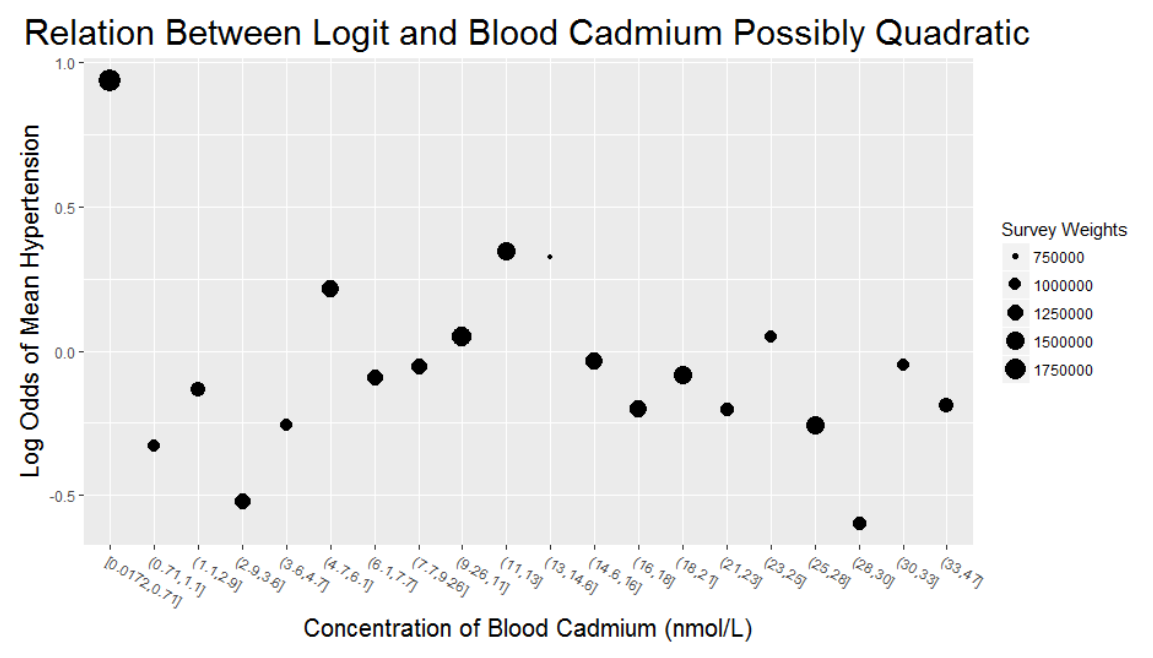
Exploratory Analysis

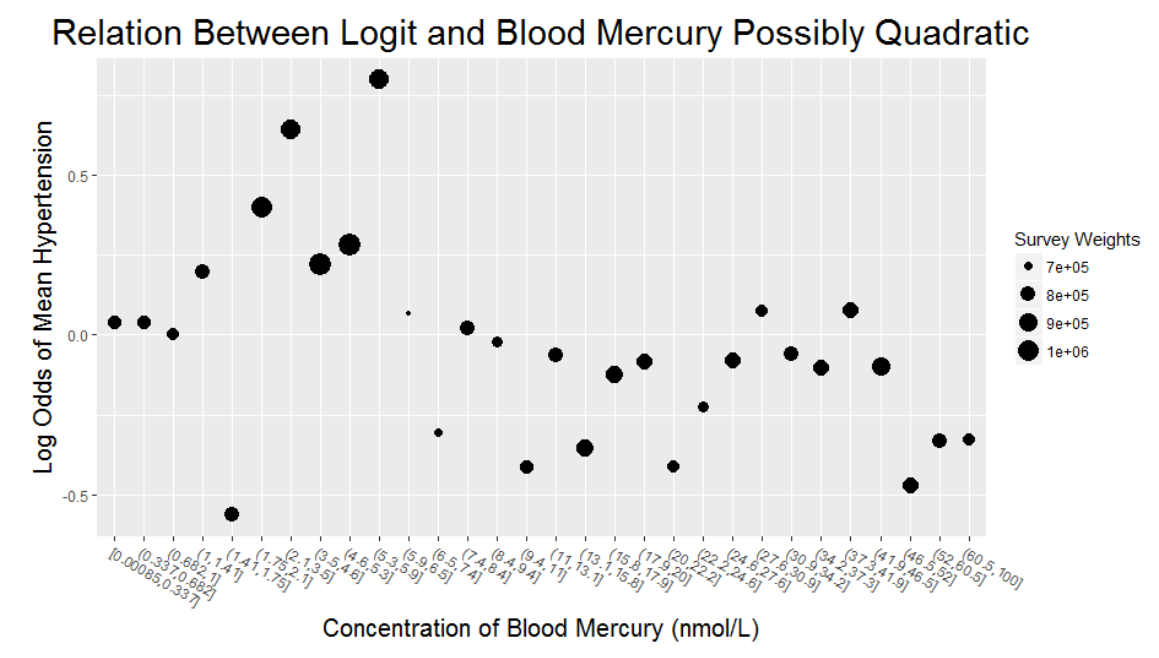
EDA: Plots were made to understand the structure of the data.

- Continuous variables were plotted as categorical to avoid observed hypertension proportions of 0 or 1.
- Since the explanatory variables exhibit low linear correlation, we can make useful plots without holding all other variables constant.









Conclusion: Including second-order terms for BMI, BCD, and BHG in the model could be informative, but isn't necessary for AGE.

Checking Model Fit

Goodness of Fit: Since the response is binary, the Hosmer-Lemeshow test statistic was used to check overall goodness of fit.

- Follows χ^2_{g-2} under null hypothesis that model fits sample data well (Hosmer et al., 1980).
- No theoretical guidance for choosing g, so we chose several values and used the average p-value.

Hypothesis Testing Theory

Null Hypothesis: Suppose β is partitioned by $(\beta_{(1)}, \beta_{(2)})$. The null hypothesis is H_0: $\beta_{(1)} = 0$, where $\beta_{(1)}$ has dimension q.

Unweighted Model: For the unweighted model, the log-likelihood ratio statistic under H_0 follows χ^2_q.

Weighted Model: The likelihood functions do not exist for the weighted model. Following Lumley & Scott, 2017, the unweighted likelihood formulas were altered to include weights:

$$\ell(\beta) = \sum_{i=1}^{n} w_i \log f(y_i|x_i;\beta) \tag{3}$$

The "working likelihood ratio test statistic" was then defined as

$$\tau = 2\left[\ell(\hat{\beta}) - \ell(\hat{\beta}^*)\right] \tag{4}$$

where $\hat{\beta}^*$ is the solution to (3) when $\beta_{(1)} = 0$. The saddlepoint approximation (Kuonen, 1999) was used to approximate the distribution of τ.

Results

Model Without Interactions: Hypothesis testing on both the unweighted and weighted model without interaction effects was carried out.

Null	p (Unweighted)	p (Weighted)	Design Effects
SMK 2 = 0 SMK 3 = 0	0.113	0.472	0.882 0.948
SEX = 0	0.001	0.005	1.14
AGE = 0	0.000	0.000	1.06
BMI = 0 $\text{BMI}^2 = 0$	0.000	0.002	1.50 1.55
BCD = 0 $\text{BCD}^2 = 0$	0.505	0.731	1.59 1.66
BHG = 0 $\text{BHG}^2 = 0$	0.048	0.185	1.38 1.43

- Very strong evidence (p-value < 0.01) that AGE , SEX, and BMI for both unweighted and weighted models affect mean HIGHBP, when other risk factors are accounted for.
- Some evidence ($0.01 <$ p-value < 0.05) that BHG affects mean HIGHBP, but only for unweighted model.
- Little to no evidence that other variables affect mean HIGHBP when other risk factors are accounted for.

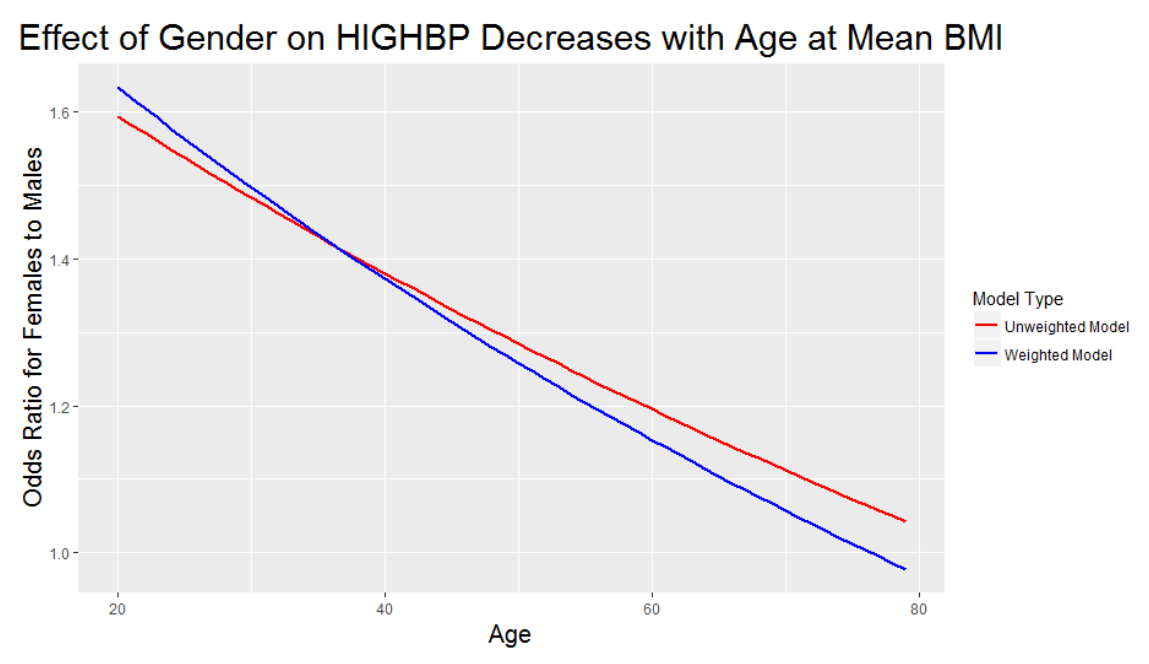
Model with Interactions: Three interaction effects are now added to both the unweighted and weighted model.

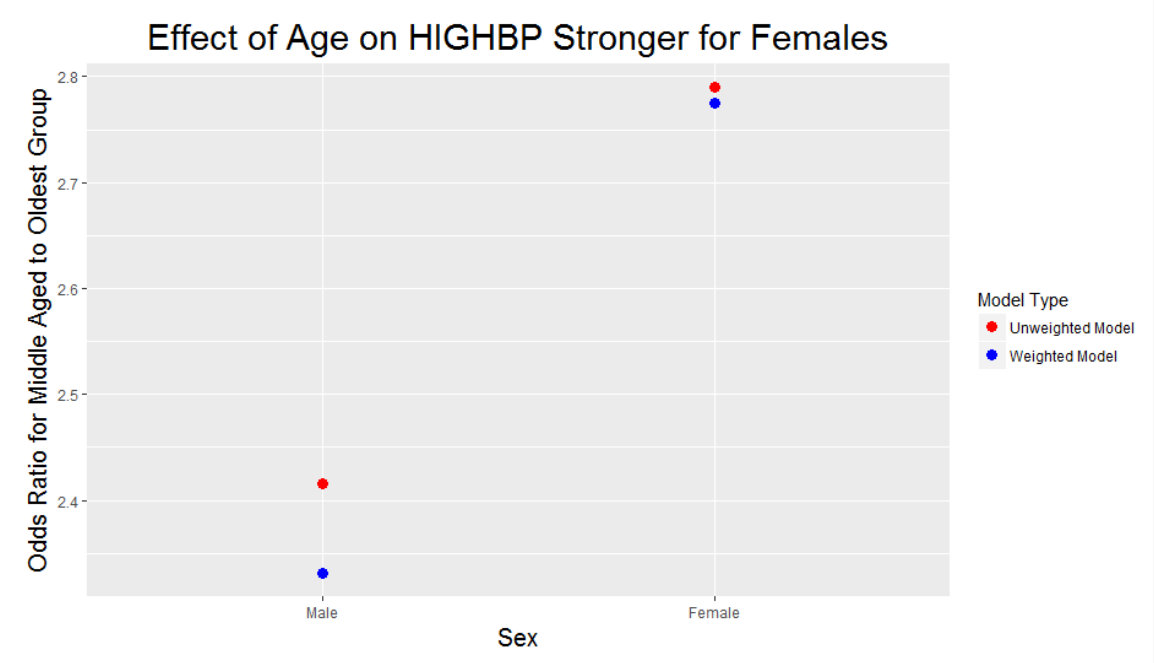
Null	p (Unweighted)	p (Weighted)	Design Effects
SEX:AGE= 0	0.143	0.048	0.869
SEX:BMI = 0	0.046	0.093	1.41
AGE:BMI = 0	0.149	0.046	1.16

- For unweighted model, some evidence (p-value < 0.05) that effect of BMI on HIGHBP depends on SEX.
- For weighted model, some evidence (p-value < 0.05) that effects of SEX and BMI on HIGHBP depend on AGE.
- Little to no evidence that other variables affect mean HIGHBP when other risk factors are accounted for.

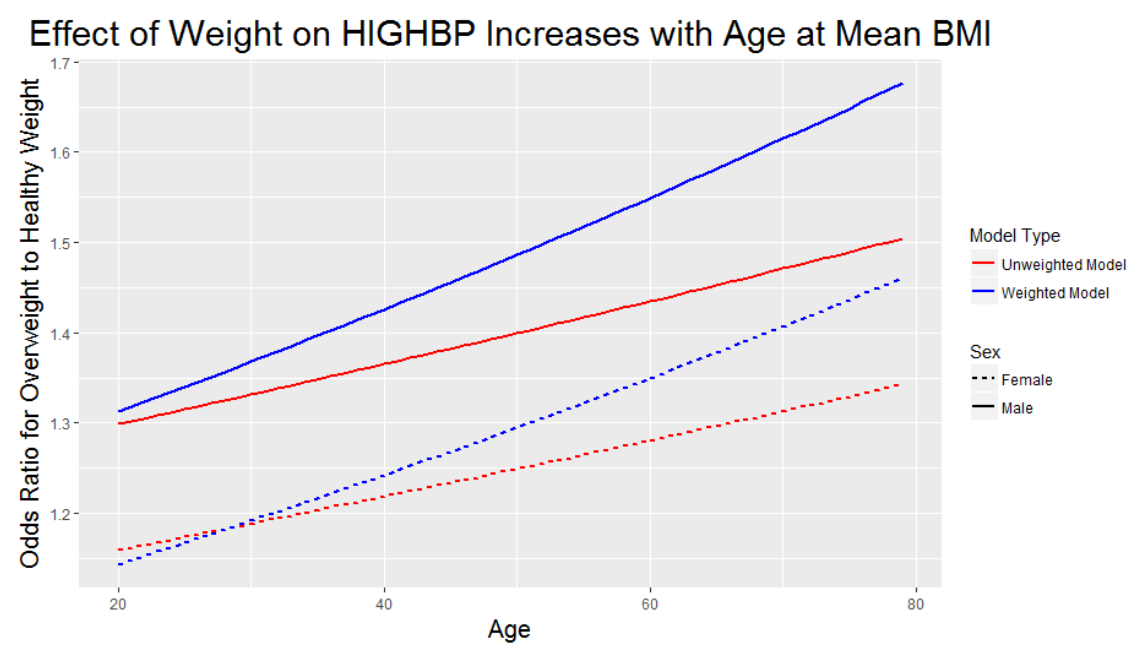
Results Continued

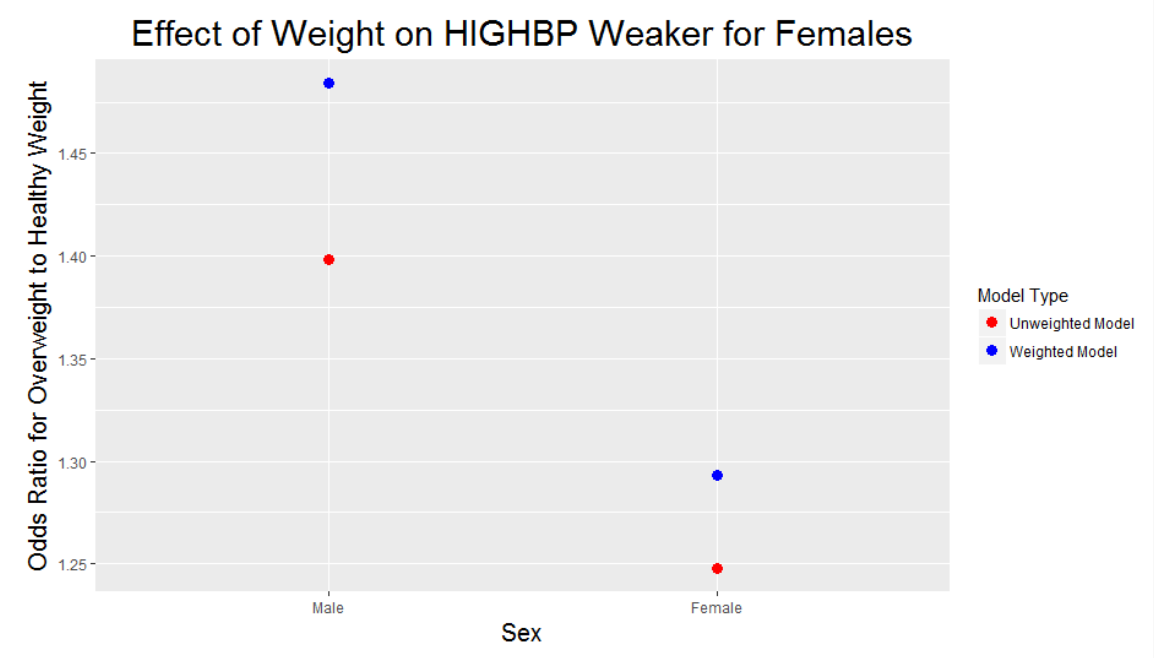
Analysis: Odds Ratio plots can clarify these interaction effects.





- At older ages, SEX is a weaker risk factor (predictor) for HIGHBP, when other variables are fixed.
- For females, AGE is a stronger risk factor for HIGHBP





- At older ages, BMI is a stronger risk factor for HIGHBP.
- For males, BMI is a stronger risk factor for HIGHBP.

Keep in mind that only for certain models (see interaction p-values) is there some evidence of these interaction effects.

Design Effects: For an estimator, the design effect is the ratio of weighted variance to unweighted variance. This makes it easy to see that BCD, BMI, BHG, and SEX:BMI have weighted variances that are quite different from their unweighted variances.

Further Work

Weighted Model Fit: The weighted model was not checked for model fit. A goodness of fit test that can take weights into account was suggested by Archer et al., 2007.

Variance of Imputation: The uncertainty in the imputed censored and missing data was ignored, and this will artificially deflate the variance of the estimators. Can use a method such as the bootstrap to estimate the error introduced by this imputation.

Improved Imputation: Use a more sophisticated method such as survival analysis to estimate the censored data.

References

- Archer, K. J., Lemeshow, S., & Hosmer, D. W. (2007). Goodness-of-fit tests for logistic regression models when data are collected using a complex sampling design. *Computational Statistics & Data Analysis*, 51(9), 4450-4464. doi:10.1016/j.csda.2006.07.006.
- Hosmer, D. W. & S. Lemeshow (1980). Goodness of fit tests for the multiple logistic model. *Communications in Statistics-Theory and Methods* A9, 1043-1069.
- Kuonen, D. (1999). *Saddlepoint approximations for distributions of quadratic forms in normal variables.* Biometrika 86 929-935.
- Lumley, T., & Scott, A. (2017). Fitting Regression Models to Survey Data. *Statistical Science*, 32(2), 265-278. doi:10.1214/16-STS605.

Acknowledgements

We are very grateful to Shaun Sun, whose guidance and technical expertise made this project possible. We also thank Thomas Lumley for making public his R survey() package that was used extensively in this analysis.