Statistical Consulting Report

Well-being of indigenous persons in Vancouver and Surrey: An analysis of the APS Survey Data

Consultants: Harper Cheng
Client: Patrick Dubois

Abstract

Complex survey data analysis requires techniques different from those used in the standard analysis or the model-based approach. To accommodate the complicated sampling process involved in data collection, the design-based approach is preferable for analyzing complex survey data. In this report, I first explain the difference between the model-based and design-based approach. Logistic and multinomial logistic regression are introduced under the model-based approach. The method is further extended to the design-based analysis where stratification and bootstrap weights are taken into consideration. A short but comprehensive tutorial on performing and interpreting logistic and multinomial logistic regression in R is provided. Also described is the use of an R package `survey` for analyzing the survey data. The complete R code is included for your reference (see Appendix for a GitHub link to the .Rmd file).

1 Introduction

To gain a better understanding of the well-being of indigenous people living in Canada, data from the Statistics Canada Aboriginal People's Survey are to be analyzed and compared across different geographical regions. Specifically, the objective is to examine the difference between aboriginal people living in Vancouver and Surrey, both in levels of reported mental health, employment status, etc, and relationships between such levels and other possible covariates (e.g. between mental health and community involvement). Due to the complex nature of the survey data, it is more appropriate to employ the design-based approach as opposed to the model-based approach used in the standard analysis. This is because the sampling process for the large-scale sample survey is not restricted to simple random sampling. Stratification, clustering, and multistage sampling could complicate the structure of survey data. To estimate parameter values and their standard errors more accurately, the design-based approach should be adopted where the survey weights, stratum/cluster variables, and bootstrap weights are incorporated into the analysis process.

Since the actual APS dataset from Statistics Canada is currently not available, a toy dataset is used for performing the analysis with the R package `survey` which implements the design-based approach to give more accurate parameter estimations.

2 Data Collection

The Aboriginal Peoples Survey (APS) is a national survey conducted by Statistics Canada on First Nations people in Canada who are older than 15 years old and are living off reserve, Métis and Inuit. It is conducted every five years and information on employment, education, health and other aspects are collected. APS design employs a two-phase design in which the first phase corresponds to the selection of eligible individuals who identified themselves as aboriginal people in the Census of Population of the previous year. The second phase is stratified sampling where geographical region, age group, and aboriginal group are used as strata.

3 Statistical Questions

This report addresses the following questions:

1. How to form prediction models for binary and ordinal responses?
2. How to analyze complex survey data?

4 Preamble

Before analyzing the survey data, it is important to understand the fundamental difference between "regular" data and survey data as well as the approach most suitable for analyzing such survey data. In addition, in preparation for the design-based approach (i.e., the analysis with survey weights and bootstrap weights incorporated), logistic and multinomial logistic regression under the model-based setting are discussed first. Once you understand the fundamentals and analytical techniques, the transition from the model-based approach to design-based approach will be smooth.

4.1 Model-based vs Design-based Approach

The standard analysis or the model-based approach cannot be applied to survey data because the sampling process for complex survey data is not necessarily simple random sampling. Stratification, clustering, and multistage sampling may result in the sample having a distribution different from the one that the model assumes. Applying standard analyses to survey data will lead to inaccurate estimations on parameter values and on their standard errors, confidence intervals, etc. Instead, the design-based approach should be adopted where the survey weights, stratum/cluster variables, and bootstrap weights are incorporated into the analysis process [3].

The survey weight makes statistics computed from the data more representative of the population. It can account for unequal selection probabilities, unit non-response, under-coverage/over-coverage and auxiliary

information obtained from other data sources. Survey weights reduce estimation bias on means and proportions [1]. However, the use of survey weights almost certainly inflates the uncertainty measures such as the standard error. This is because by adjusting the above-mentioned issues with survey weights, additional source of variation is introduced [2]. To provide interval estimation of the parameter value, the bootstrap method is often adopted where subsamples of the original sample space are drawn with replacement in a repeated manner. Estimates of the parameters can then be calculated based on the empirical distribution of the parameter (more on this topic in the later section). Thus, the survey weights are used to estimate a parameter value while bootstrap weights are used for estimating the variance of the parameter.

Unlike regular bootstrap method where the subsamples are chosen randomly with equal probability, for complex survey data, extra information is needed for correctly choosing the bootstrap samples. Statistics Canada uses the rescaling bootstrap or Rao-Wu bootstrap which is a special bootstrap method used for stratified multi-stage designs. However, the rescaling method is not easy to carry out because of the amount of math involved and extra information needed. Instead of letting users do the rescaling bootstrap themselves, Statistics Canada uses bootstrap weights as an improved technique derived from the rescaling bootstrap [4]. The rescaling bootstrap and bootstrap weights achieve the same results, but the bootstrap weights are more user-friendly. Now that we have the bootstrap weights, the regression model can be fitted repeatedly to the data with a different set of bootstrap weights for each round. The mean estimated parameter value and the corresponding standard error and confidence interval can then be obtained from the empirical distribution of the estimated parameter.

4.2 Logistic Regression and Multinomial Logistic Regression

This subsection is an introduction to the logistic and multinomial logistic regression **under the model-based design setting** (i.e., no survey weights or bootstrap weights are incorporated). The framework can be easily extended to the design-based approach.

4.2.1 Logistic Regression

A linear regression is used for modelling continuous and normally distributed responses. When a response fails to meet the criteria required for a linear regression, the *generalized linear model (GLM)* serves as an excellent extension to the linear regression model. It can be used to model responses that follow distributions in the exponential family which includes normal, binomial, Poisson, etc. That is to say, the response is not limited to continuous; discrete and count data are allowed in GLM as well.

Specifically, the logistic regression is commonly used to model the binary or binomial data. For the purpose of this report, I will use binary data as an example. Formally, let the response be Y where it takes on two values, 0 and 1, and the probability of Y being 1 given the explanatory variables x is denoted as $P(Y = 1|x)$. Since for a binary response, the response is either 1 or 0, the probability of Y being 0 is

therefore $P(Y=0|x) = 1 - P(Y=1|x)$. The mean of the response is calculated as

$$E(Y|x) = 1 \times P(Y=1|x) + 0 \times P(Y=0|x) = P(Y=1|x).$$

Thus, the probability of Y being 1 is also the mean of the binary response. Suppose there are n number of explanatory variables, $x = \{x_1, \ldots, x_n\}$, with corresponding regression coefficients being $\beta = \{\beta_0, \beta_1, \ldots, \beta_n\}$ where β_0 denotes the y-intercept. The logistic regression can be expressed as follows:

$$\log\left(\frac{P(Y=1|x)}{1-P(Y=1|x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \tag{1}$$

where the term $\frac{P(Y=1|x)}{1-P(Y=1|x)}$ is called the *odds* and it measures the relative probability of the event happening to the probability of the event not happening. The logarithm of odds, $\log\left(\frac{P(Y=1|x)}{1-P(Y=1|x)}\right)$, is called *log-odds* and it is commonly referred to as the *logit function* (thus the name logistic regression). The reason for introducing the logit function is that the range of the mean response, i.e., $E(Y|x) = P(Y=1|x)$, can only take on values from 0 to 1. However, the linear combination of the explanatory variables $\beta_0+\beta_1 x_1+\cdots+\beta_n x_n$ or the so-called *linear predictor* takes on any value from -∞ to ∞. The logit function maps the range of the response to a real line such that it matches the domain of the linear predictor.

The interpretation of the logistic regression is comparable to that of the linear regression, except the mean of the response is now replaced with the log-odds of the mean response. In other words, 1 unit increase in x_1 while keeping everything else constant would result in β_1 amount of change in the log-odds of the mean response. To interpret the result on the original scale, we can simply exponentiate both sides of Equation 1, that is

$$\text{odds} = \frac{P(Y=1|x)}{1-P(Y=1|x)} = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n).$$

If x_1 increases by 1 unit while keeping other explanatory variables constant, then we have

$$\begin{aligned}
\text{odds}_{\text{before}} &= \exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n\} \\
\text{odds}_{\text{after}} &= \exp\{\beta_0 + \beta_1 (x_1+1) + \cdots + \beta_n x_n\} \\
\frac{\text{odds}_{\text{after}}}{\text{odds}_{\text{before}}} &= \exp(\beta_1).
\end{aligned}$$

If x_1 is increased by 1 unit, the estimated odds would change by a factor of e^{β_1}. In other words, we would expect a percentage increase of $e^{\beta_1} - 1$ in the odds of Y being 1 if β_1 is greater than 0; or a percentage decrease of $1 - e^{\beta_1}$ if β_1 is less than 0. That is to say, $1 - e^{\beta_1}$ or $e^{\beta_1} - 1$ can be interpreted as the percentage change in the odds of Y being 1 with 1 unit increase of x_1. If you are more interested in the mean response or the probability of Y being 1 given a set of values on the explanatory variables, with some mathematical manipulations, it can be calculated as

$$P(Y=1) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}.$$

Here is an example of performing logistic regression in R. The binary response is taken to be the employment status where 1 represents employed, 0 represents unemployed and NA otherwise. The covariates are chosen to be age and gender.

```
> logistic.fit <- glm(DEMPSTAT~AGE_YRSG+SEX, data=dat, family=binomial)
> summary(logistic.fit)

Call:
glm(formula = DEMPSTAT ~ AGE_YRSG + SEX, family = binomial, data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6685  -1.0205   0.7889   0.9806   1.4022

Coefficients:
                                     Estimate Std. Error z value Pr(>|z|)
(Intercept)                          -0.41588    0.05099  -8.156 3.47e-16 ***
AGE_YRSGBetween the ages of 19 and 24  0.89823    0.05717  15.713  < 2e-16 ***
AGE_YRSGBetween the ages of 25 and 34  1.26642    0.06358  19.919  < 2e-16 ***
AGE_YRSGBetween the ages of 35 and 44  1.52214    0.06702  22.713  < 2e-16 ***
AGE_YRSGBetween the ages of 45 and 54  1.36159    0.06449  21.113  < 2e-16 ***
AGE_YRSGAge 55 and over                0.03481    0.05519   0.631 0.528263
SEXFemale                            -0.09853    0.02929  -3.364 0.000768 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Because age and gender are converted to factors, R adopts dummy coding where the baseline level for age is chosen to be "Between the ages of 15 and 18" and the baseline for gender is chosen to be "male". Interpretationwise, the odds of being employed for female is $1 - e^{-0.09853} = 9.4\%$ lower compared to their male counterparts in the same age group.

4.2.2 Multinomial Logistic Regression

A generalization of the logistic regression is multinomial logistic regression. It allows for ordinal responses or, more generally speaking, categorical responses that have more than two levels. Let us build from the setup in the previous section where the response Y now takes on K levels. Instead of modelling $P(Y = 1|x)$ as we did in the logistic regression, now that there are K levels, we are interested in modelling $P(Y = 1|x), \ldots, P(Y = K - 1|x)$ with respect to the baseline level $P(Y = K|x)$, that is, we model the log-odds of each level relative to the baseline level. Different choices of the baseline level would affect the estimated parameter values, but the resulting probabilities and interpretations are invariant under different choices of the baseline level. The explanatory variables are $x = \{x_1, \ldots, x_n\}$ with corresponding regression coefficients

$\beta_k = \{\beta_{0,k}, \beta_{1,k}, \ldots, \beta_{n,k}\}$ for the k-th level where $k = 1, \ldots, K-1$. In other words, there are now $K-1$ sets of separate regression coefficients for each level (excluding the baseline level). The framework of multinomial logistic regression is as follows:

$$\begin{aligned}
\log\left(\frac{P(Y=1|x)}{P(Y=K|x)}\right) &= \beta_{0,1} + \beta_{1,1}x_1 + \cdots + \beta_{n,1}x_n \\
\log\left(\frac{P(Y=2|x)}{P(Y=K|x)}\right) &= \beta_{0,2} + \beta_{1,2}x_1 + \cdots + \beta_{n,2}x_n \\
&\cdots \\
\log\left(\frac{P(Y=K-1|x)}{P(Y=K|x)}\right) &= \beta_{0,K-1} + \beta_{1,K-1}x_1 + \cdots + \beta_{n,K-1}x_n
\end{aligned} \tag{2}$$

Since all probabilities from K levels sum up to 1, the probability of the baseline level can be easily derived as

$$P(Y=K|x) = \frac{1}{1 + \sum_{i=1}^{K-1} \exp\left(\beta_{0,i} + \beta_{1,i}x_1 + \cdots + \beta_{n,i}x_n\right)}.$$

The probabilities for other levels can be obtained accordingly:

$$P(Y=k|x) = \frac{\exp\left(\beta_{0,k} + \beta_{1,k}x_1 + \cdots + \beta_{n,k}x_n\right)}{1 + \sum_{i=1}^{K-1} \exp\left(\beta_{0,i} + \beta_{1,i}x_1 + \cdots + \beta_{n,i}x_n\right)}, \quad k = 1, \ldots, K-1.$$

Implementation and interpretation are demonstrated with the following R example where the ordinal response `DLASTWKG` describing the number of months since someone last worked is regressed against the covariate gender.

```
> multinom.fit <- multinom(DLASTWKG~SEX, data=dat)
> summary(multinom.fit)

Call:
multinom(formula = DLASTWKG ~ SEX, data = dat)

Coefficients:
  (Intercept) SEXFemale
2  -0.5233744 0.2136986
3  -1.0082271 0.4758590
4   0.2059472 0.6139619
```

The response is recoded to take on levels from 1 to 4 with increasing time of being unemployed. The baseline level for the response is taken to be 1 where it encodes "Has not been working for less than six months". We will compare response level 2 ("Has not been working for six months to less than one year") to the baseline level, that is

$$\log\left(\frac{P(Y=2|x)}{P(Y=1|x)}\right) = -0.523 + 0.214x_{\text{Female}}$$

where $x_{\text{Female}} = 1$ for female and 0 for male. Thus, the odds of the response being 2 (i.e., being unemployed

longer) compared to the baseline is $e^{0.214} - 1 = 23.9\%$ higher for female than male.

5 A Working Example of Analyzing Survey Data with `survey`

The R package `survey` is designed to facilitate the analysis of complex survey data where it provides the option for accommodating sampling designs more complicated than simple random sampling. The main feature of the package is that we need to specify a "design object" prior to mathematical manipulations and model fitting. A design object functions as a vessel containing all information about the sampling design and it is used in lieu of the actual dataset such that the characteristics of the sampling design (e.g., stratification for APS data) can be taken into account during the analytical process. The function `svydesign()` is used to create the design object. You can find more information about this function in the package manual (here is the link: `https://cran.r-project.org/web/packages/survey/survey.pdf`).

In this section, I provide some snippets of R code to demonstrate the data analysis using the `survey` package. A link to the complete R code can be found in the Appendix. Please note that the specification of the sampling design may not be reflective of the actual APS design. It is important to understand the sampling design before conducting any analysis to ensure the validity of the derived statistical inference. I suggest you consult with the researcher you are working with to gain a better understanding of the APS design.

5.1 Exploratory Analysis

As mentioned in 4.1 Model-based vs Design-based Approach, the use of survey weights is enough to yield accurate estimations of means and proportions which are the summary statistics of interest for exploratory analyses. The design object can be specified as

```
> sw.des <- svydesign(data=dat, weights=~PUMFWGHT, id=~PUMFID, strata=~AGE_YRSG)
```

where the survey weight is used as the sampling weight, individual ID is used to identify the primary sampling unit, and age groups serve as the strata. Since most of the response variables in the APS dataset are categorical, proportions can be calculated and visualized in a side-by-side bar plot. The function `svytable()` returns crosstabulations of the response and covarites. For example, the code below returns a 2 by 2 table of employment status and gender.

```
> tbl <- svytable(~DEMPSTAT+SEX, design=sw.des)
> summary(tbl)

        SEX
DEMPSTAT   Male Female
       0 182972 228899
```

```
       1 284491 301369

Pearson's X^2: Rao & Scott adjustment

data:  svychisq(~DEMPSTAT + SEX, design = sw.des, statistic = "F")
F = 16.531, ndf = 1, ddf = 20843, p-value = 4.804e-05
```

If you call the table by `summary()`, it performs a Chi-square test for testing independence between these two variables. A p-value smaller than the significance level 0.05 provides evidence on rejection of the null hypothesis (two variables are independent) which means that employment status is dependent of gender. Unfortunately, `svytable()` does not support calculations of confidence intervals making it difficult to include error bars on the bar plot as you requested. However, even if there is a way of obtaining the confidence intervals, they are calculated under the sampling design where survey weights are used. As discussed earlier, the use of survey weights alone results in large standard errors and thus wider confidence intervals. Due to the inaccuracy of estimated errors, I suggest plotting the bar plot without error bars. The standard error of an estimate can be tended to in the confirmatory analysis.

5.2 Confirmatory Analysis: Regression

When performing a regression analysis, not only is the estimated regression coefficient of interest, but the standard error is an important part of the statistical inference as well. The standard error assesses how accurate an estimate is compared to its true value, that is, the smaller the standard error, the smaller the uncertainty. Thus, for regression analyses, bootstrap weights are used in place of survey weights for more accurate parameter estimations (i.e., smaller standard error).

5.2.1 Logistic Regression

To properly utilize bootstrap weights, suppose we are interested in estimating regression coefficient β, and the APS dataset comes with B bootstrap weights. Each time, a logistic regression is fitted to the data using one set of bootstrap weights resulting in an estimate $\hat{\beta}_b$, $b = 1, \ldots, B$ where the hat symbol $\hat{}$ represents the estimated value of a parameter. The idea is to fit the model repeatedly, each time using a different set of bootstrap weights. The end result is B sets of estimated coefficient, $\hat{\beta}_1, \ldots, \hat{\beta}_B$, giving rise to an empirical distribution of $\hat{\beta}$. The mean and variance of $\hat{\beta}$ are then calculated from the empirical distribution [2]

$$E(\hat{\beta}) = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_b, \quad Var(\hat{\beta}) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\beta}_b - E(\hat{\beta})\right). \tag{3}$$

In terms of implementation in R, a "for loop" is a great way of achieving the above repetition. For each iteration, we first create a design object with one set of bootstrap weights. The object is then used for carrying out the regression using `svyglm()` with the `family` argument specified to be `binomial` for logistic regression. The resulting parameter estimates from each iteration are stored in a data frame for calculations of the mean,

standard error, and confidence interval. On a side note, it is also possible to use `glm()` function with an extra argument `weights=` to accommodate the bootstrap weights. The regression coefficient estimates given by `svyglm()` and `glm()` with `weights` are the same. However, `svyglm()` handles the bootstrap weights in a more proper manner which results in a more reliable and accurate estimation of the standard error.

In the .Rmd file, I also provide a quick sanity check and show that the standard errors calculated from the analysis with bootstrap weights are smaller than those obtained from using survey weights. This observation verifies previous statements that analyses with bootstrap weights yield more accurate parameter estimates.

5.2.2 Multinomial Logistic Regression

The `survey` package has yet to develop a function to implement the multinomial logistic regression. The lack of this functionality poses some technical difficulty when it comes to analyzing categorical responses. I have found two alternative options.

The first one is to switch gears and fit the multinomial logistic regression using standard analyses (i.e., using `multinom()` function from the `nnet` package) with an extra argument `weights` specifying the bootstrap weights.

```
> survey_multi1 <- multinom(DLASTWKG~SEX, data=dat, weights=WRPP0001)
> summary(survey_multi1)
```

Estimations of the parameter coefficients can be obtained from the empirical distribution as described in the previous section. The advantage of this approach is that the output is easy to interpret. However, the sampling design such as stratification is not being considered in the standard analysis. The validity of the obtained inference might be questionable.

The second option is to use `svyolr()` from the `survey` package to fit a *proportional odds model*. Unlike the multinomial logistic regression discussed previously, the proportional odds model models the log-odds of the response Y being less than or equal to a particular category, that is

$$\log\left(\frac{P(Y \leq j|x)}{P(Y > j|x)}\right) = \beta_{0,j} + \beta_1 x_1 + \cdots + \beta_n x_n, \quad j = 1, \ldots, J-1 \tag{4}$$

where each level j has an intercept of its own. It can be easily shown that the probability of Y being less than or equal to category j is

$$P(Y \leq j|x) = \frac{\exp(\beta_{0,j} + \beta_1 x_1 + \cdots + \beta_n x_n)}{1 + \exp(\beta_{0,j} + \beta_1 x_1 + \cdots + \beta_n x_n)} \tag{5}$$

and the probability of Y for being in category j is therefore

$$P(Y = j|x) = P(Y \leq j|x) - P(Y \leq j-1|x). \tag{6}$$

Here is an example where the ordinal response `DLASTWKG` and covariate `SEX` are used to fit a proportional odds model.

```
> des2 <- svydesign(data=dat, weights=~WRPP0001, id=~PUMFID, strata=~AGE_YRSG)
> survey_multi2 <- svyolr(as.factor(DLASTWKG)~SEX, design=des2)
> summary(survey_multi2)

Call:
svyolr(as.factor(DLASTWKG) ~ SEX, design = des2)

Coefficients:
             Value Std. Error  t value
SEXFemale 0.487568 0.06890835 7.075602

Intercepts:
    Value    Std. Error t value
1|2  -0.8927   0.0582   -15.3317
2|3  -0.1393   0.0539    -2.5860
3|4   0.3468   0.0534     6.4953
```

Remember that response level 1 encodes "Has not been working for less than six months", and response level 2 encodes "Has not been working for six months to less than one year". Take the second line of the "Intercepts" chunk for example, we can write out the expression for the log-odds of someone who has been unemployed for less than one year, that is

$$\log\left(\frac{P(Y \leq 2)}{P(Y > 2)}\right) = -0.1393 + 0.487568 x_{\text{Female}}.$$

The probability of interest can then be calculated using Equation 5 and Equation 6.

The proportional odds model requires one extra step which is to check on the proportional odds assumption. As can be seen from Equation 4, the model assumes that the only parameter that varies between each pair of outcome levels is the intercept while the slope parameters are the same for all pairs. But the data do not necessarily follow such a pattern and a violation to the model assumption would render the model invalid. For the above example, the parameter coefficients can be computed and compared as follows:

```
> ex1 <- svyglm(I(DLASTWKG>1)~SEX, des2, family="binomial")
> ex2 <- svyglm(I(DLASTWKG>2)~SEX, des2, family="binomial")
> ex3 <- svyglm(I(DLASTWKG>3)~SEX, des2, family="binomial")
> print(c(coef(ex1)[2], coef(ex2)[2], coef(ex3)[2]))

SEXFemale SEXFemale SEXFemale
0.5595928 0.5162177 0.4372366
```

There is some discrepancy across levels, but the difference is not severely large. Thus, the proportional odds model might be appropriate for modelling the given data. If many parameter coefficients are to be compared, a line graph can be used where estimated coefficients obtained from each level are plotted and compared. Details can be found in Further Readings item 4 (Proportional odds model assumption). If a clear violation of the proportional odds assumption is observed, then you might consider switching back to the model-based approach as discussed earlier.

6 Conclusion

This report briefly discusses the difference between the model-based and design-based approach including the importance of adopting the design-based approach for analyzing complex survey data. Also discussed are introduction, interpretation, and implementation of the logistic regression and multinomial logistic regression under the model-based approach setting. A short tutorial is also provided on the use of `survey` package where the bootstrap weights are incorporated into the analysis process. In particular, when modelling the ordinal response, the proportional odds model is recommended provided that the assumption is met. Otherwise, multinomial logistic regression with bootstrap weights can be used as an alternative. But this option does not accommodate specifications on the sampling design and might lead to inaccurate inference. The most important suggestion is that prior to data analysis, it is advised to consult with your colleagues who are more familiar with the APS sampling design in order to correctly specify the design object for the survey analysis.

7 Further Readings

1. A seminar on the difference between model-based and design-based approach:
   `https://crdcn.org/sites/default/files/cool_feb2011_stc_template_final.pdf`
2. Multinomial logistic regression analysis and interpretation in R:
   `https://stats.idre.ucla.edu/r/dae/multinomial-logistic-regression/`
3. `survey` manual: `https://cran.r-project.org/web/packages/survey/survey.pdf`
4. A more detailed tutorial on survey data analysis with `survey`:
   `https://stats.idre.ucla.edu/r/seminars/survey-data-analysis-with-r/`
5. Proportional odds model interpretation:
   `https://data.library.virginia.edu/fitting-and-interpreting-a-proportional-odds-model/`
   `https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/`
6. Proportional odds model assumption:
   `https://rpubs.com/corey_sparks/58926`

Appendix

Complete R code for data processing and data analysis can be accessed from my Github `https://github.com/harpercheng91/StatsConsulting/blob/main/survey.Rmd`.

References

[1] Lahiri, P. (2003). On the impact of bootstrap in survey sampling and small-area estimation. *Statistical Science*, 18(2):199–210.

[2] Lenka Mach, A. S. and Pettapiece, R. (2007). Study of the properties of the rao-wu bootstrap variance estimator: What happens when assumptions do not hold? SSC Annual Meeting.

[3] Roberts, G. (2011). Some considerations when choosing analytical methods and tools for complex survey data.

[4] Zeinab Mashreghi, D. H. and Léger, C. (2016). A survey of bootstrap methods in finite population sampling. *Statistics Surveys*, 1(52).