Statistical challenges in observational cohort studies

Hof, M.H.P.

Publication date
2015
Document Version
Final published version

8 Conclusions and further research

In this thesis we have investigated different topics related to observational cohort studies, both methodological and applied. In the first part we proposed a new method for sampling individuals from a finite population when we have heterogeneous willingness to participate in a cohort study. The second part of this thesis was focused on developing new methods to analyze data obtained with record linkage. In the third part of this thesis we proposed a quasi-Monte Carlo approach to solve the computational problems that arise in joint modeling of multiple longitudinal markers, multiple recurrent events, and a terminal event. Finally, in the fourth part we investigated two aspects of childhood growth with data from cohort studies. We measured differences between growth of infants with a native and immigrant background and estimated the association between the peak in the body mass index (BMI) around nine months of age and later body composition measures and blood pressure. For all parts of this thesis, conclusions and directions for further research are now given.

Part I: Sampling

In chapter 2, we extended the list sequential sampling method, developed by Bondesson and Thorburn [23], to deal with unknown heterogeneous willingness to participate in a cohort study and with delayed response to the invitation. When characteristics that are related to the willingness to participate are known for each individual from the population, our adaptive list sequential sampling method alters the invitation probabilities to deal with the heterogeneous willingness to participate. To estimate the probability to participate, the adaptive list sequential sampling method uses the response to the invitation of previously invited individuals. In addition, delayed response to the invitation during the recruitment period is successfully dealt with by using an estimated probability to participate when

an individual has not responded yet. Our simulations showed that the adaptive list sequential sampling method can recruit samples with some desired composition, and the simulations also showed that the adaptive list sequential sampling method is robust to misspecification of the participation probability model.

Although the adaptive list sequential sampling method allows us to deal with heterogeneous willingness of individuals to participate in the study and their delayed response, in real life situations we are usually confronted with extra challenges. We developed our method by assuming that we only need to evaluate once if an individual from the population should be invited or not. Usually, individuals who do not respond to the invitation are re-invited. Our adaptive list sequential sampling method can be extended to sampling designs in which we use multiple invitation rounds. In addition, as suggested in the discussion of chapter 2 (see page 30) different invitation techniques are frequently used in cohort studies to approach individuals from the population. Each invitation technique usually has a different rate of success and, generally, higher participation probabilities can be expected with face-to-face recruitment of participants compared to methods that are less personal such as telephone contact or invitation by letter [57]. Moreover, the more personal the recruitment method becomes, the more resources it requires. Therefore, some trade-off between costs and participation is necessary to obtain samples with some desired composition. Further research could be focused on how to include this trade-off in our newly proposed method.

Part II: Record linkage

In chapters 3 and 4 we proposed new methods to analyze data obtained from record linkage. In chapter 3, a weighted least squares (WLS) approach was proposed for the analysis of data obtained from combining two datasets with a record linkage strategy. Our WLS method can be used for generalized regression models. We showed with simulations that, when the linkage variables have sufficient discriminative properties (i.e. a large number of unique values and few typing errors), the WLS method can produce accurate results.

The main disadvantage of the WLS method is that it only uses the linkage variables to determine whether a record pair from two datasets are a match. However, the covariates and the outcome of the regression model might contain additional information on whether a record pair is a match. Therefore, in chapter 4, we

proposed a mixture model for data obtained with record linkage that used the information from all observed data. With simulations, we showed that the mixture model gives accurate parameter estimates in many situations and, most importantly, also in some situations when WLS fails.

Further research in developing methods to analyze data obtained with record linkage should be focused on extending the mixture model from chapter 4. The mixture model could be extended to fit hierarchical data such as family data (i.e. correlation between records containing the covariates measurements in dataset **A**) or repeated measures data (i.e. correlation between records containing the outcome measurements in dataset **B**). Another interesting extension is to adapt other constraints in our model regarding the matching status. On page 65, we showed that a situation in which each record from **A** and **B** has *at most* one match in the other dataset gives a marginal likelihood which is too computationally expensive to calculate. In chapter 4, we reduced the computational burden by assuming independence between *all* record pairs and showed that this did not substantially reduced the performance of the model. However, an alternative is to assume that each record from **A** has *at most* one match in **B**; instead of assuming independence between all record pairs, we assume independence between 'blocks' of record pairs. However, whether this increases the accuracy of the mixture model or whether this is computationally feasible requires more research.

Part III: Joint modeling

In chapter 5 we proposed an estimator based on quasi-Monte Carlo (QMC) approximation to estimate the parameters of joint models with a large number of markers and (recurrent) event types. Our simulations showed that, although the coverages of the 95% confidence intervals of the parameters regarding the markers were too low, the estimator based on QMC approximation gave accurate estimates of the parameters regarding the event types and also accurate estimates of the associations between the markers and event types. Further research could be focused on developing methods to determine whether the number of QMC vectors from the random effects and frailty distributions is sufficiently large to obtain accurate results. A starting point might be the statistical test proposed by Hajivassiliou [132]. Another area for further research is to develop semi-parametric estimators which use the QMC approximation technique, in which we make no assumptions about the nature or shape of the baseline hazard functions regarding the recurrent events and the terminal event.

Part IV: Childhood growth

In chapter 6 we confirmed differences in growth between children with a Dutch background and children with a Moroccan or Turkish background. On average, between the ages 0-5 years, the weight of Moroccan and Turkish children increases faster than for the average Dutch child. In addition, we also found a difference in the growth pattern of children with a Dutch background and children with a Surinamese background. The average Surinamese child tends to be lighter than the average Dutch child between the ages 0-5 years. Further research is necessary to find causes of differences between children with a native Dutch background and children with an immigrant background. In addition, because the largest differences were found in the weight for age curves, more research is needed to rule out that the observed differences were caused by different feeding behaviors among the ethnic groups. Since nutrition is one of the most important determinants of growth [206, 207], it has a large impact on the growth of the child [208]. When the differences between the ethnic groups can be explained by different feeding patterns, the necessity of ethnic specific growth charts should be re-evaluated. A risk of having ethnic specific growth charts in this situation is that we not correct for ethnic differences in growth but for unhealthy feeding patterns. Unfortunately, the feeding behaviors were not registered accurately enough to investigate this in more detail.

In chapter 7 we found significant associations between the age and height of the BMI peak around nine months of age and anthropomorphic measures at 5-6 years of age. The results showed that a higher and later BMI peak resulted in higher BMI, higher fat percentage, and higher waist to height ratio in boys and girls at the age of 5-6 years. No associations were found between age and height of the BMI peak and later blood pressure. In addition, we showed that the BMI at nine months of age could be used as a proxy of the BMI peak. Further research is necessary to determine whether we can predict the age and height of the BMI peak with pre-pregnancy mother characteristics and characteristics of the child at birth. Moreover, the nutrition that is given to the child between 0-9 months of age could be of interest since it might influence the weight increase of the child. If we can determine which children are at risk of a later and higher BMI peak, targeted nutritional advise could be given to their mothers.

Another interesting topic for further research is to investigate whether the characteristics of the BMI peak are predictive for anthropomorphic measures in (early) adulthood. Although there have been analyses performed to measure the associ-

ations between characteristics of the adiposity rebound and health measures in adulthood [49, 50, 48, 182], this has not been done for the BMI peak. Moreover, other interesting analyses that might be performed are to measure the correlation between the BMI peak and the adiposity rebound, or to investigate whether a combination of the BMI peak and the adiposity rebound is predictive for later health status.