Cautions to Be Taken for Parameter Estimation from Follow-up Study with Loss of Follow-up

Abstract

The validity and precision of one's estimation are usually two essential concerns in most of the descriptive studies. In this report, I would like to especially dig into these two concerns in a descriptive "follow-up" study, a study approach by which one can estimate a population parameter such as incidence rate, which is a common rate of interest in epidemiology. The reason why I hope to look more into the estimation issue in a follow-up study is because at least from my own intuitive opinion, the "loss of follow-up" problem, which might happen under the study period of a follow-up study, could pose a potential threat to the final estimation process with the impact that might be somehow understated in the textbook written by Prof. Wang. To show dissatisfaction and disagree with the Prof. Wang's opinion toward the loss of follow-up problem, I pictured the variation of validity and precision of estimation from studies under different conditions through a simulating approach. I manipulated three variables: "population parameter", "proportion of loss of follow-up" and "level of association between loss of follow-up and the event of interest" (This variable will be abbreviated as "level of association" below) in the hope of trying to convince Prof. Wang that the loss of follow-up might come up as a serious problem beyond his imagination at least under some conditions (some combinations of the three variables). Finally, my simulation result shows that although under most of the conditions, the adverse effect of loss of follow-up does not seem to matter a lot, cautions should be especially taken when our sample size is relatively "large", which might not be so intuitive since large sample size might usually be considered "safer" when performing estimation.

Introduction

One kind of study in epidemiology is called "descriptive study", in which the estimation for population parameters is usually essential, by which one "describes" his/her population of concern. Statistically, one can perform this kind of estimation through his/her sample statistics with the assumption that the sample at hand is representative enough. However, it is usually neither an easy task for an investigator to decide whether his/her sample is representative enough nor how confident he/she can be to say about the representativeness.

There are several sentences from textbook "Basic Principles and Practical Applications of Epidemiological Research" written by Prof. Wang state that if there is a high proportion, say > 20% of loss of follow-up, then one should try to determine

the reasons for the loss. I am not satisfied with this kind of statement. First, I think there should be several factors other than the proportion of loss of follow-up to be considered when one is not sure whether the loss of follow up would matter a lot on his/her estimation. For instance, I think one should also take the population parameter (real parameter) and the level of association into account before trying to say anything about the validity and precision. Before talking about the reason why I think the two factors above should also be considered, I would like to briefly explain what the "level of association" means in this report. It is the abbreviation of "level of association between the loss of follow-up and the event of interest", which is defined as the proportion of the events of interest that are mistaken for the loss of follow-up. Thus, the higher the level of association is, the more the mistakes might happen under one's study.

The reason why the population parameter might also affect one's estimation is because the population parameter might be associated with the incident rate (occurring of the event of interest) of the loss of follow-up given the level of association. For example, as the population parameter comes higher, the more events of interest might occur, so are the events that might be mistaken for the loss of follow-up, given a constant level of association, which means that one might then come up with a biased estimation. As for the reason why the level of association might affect one's estimation is because the level of association directly decides the proportion of the events of interest that might be mistaken for the loss of follow-up. For example, now given that the population parameter is constant, as the level of association comes higher, which means the higher the proportion of events of interest might be mistaken for the loss of follow-up, it should then become more likely for one to obtain a biased estimation.

The second point that I am not satisfied with the statement in Prof. Wang's textbook is that he did not mention the "sample size" effect which might potentially cause adverse impact on one's estimation. In general, one might think that it is always good for someone to perform estimation based on a large sample size, and thus most of the investigators might be encouraged to have their sample size as large as possible in the hope of obtaining "good" estimation with high validity (low bias) and high precision (stability). However, one should take care of the fact that large sample size makes "good" estimation is only limited to the conditions where the assumptions of the "central limit theorem" are well satisfied. The "independence assumption" is one of the assumptions that I suspect to be unsatisfied under the topic of this report (follow-up study with loss of follow-up). Intuitively, since there might be association between loss of follow-up and the event of interest, I think this kind of association might somehow impair the independence of one's sampling, and thus the central limit

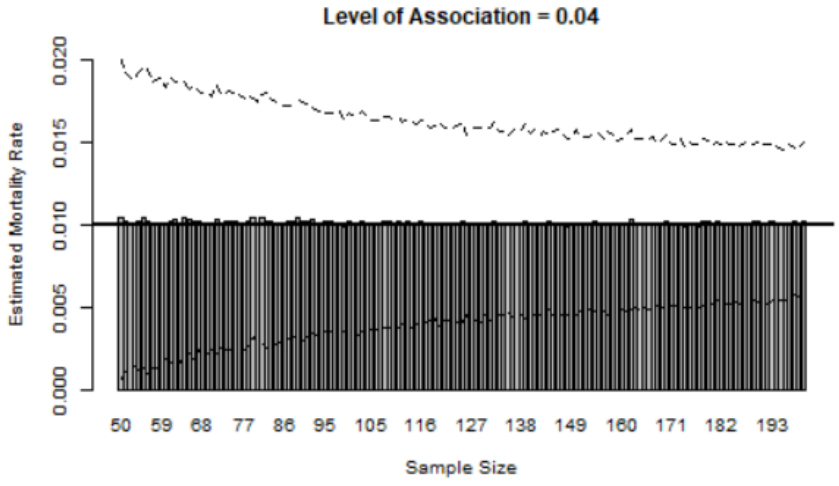
theorem might not be able to applied under follow-up study with “loss of follow-up”, which means larger sample size might not always guarantee a “good” estimation.

Method

To investigate the separate main effects and interactions among the three variables: “population parameter”, “proportion of loss of follow-up” and “level of association”, I have done a simulation research to try to picture the patterns of the change of the validity and precision along each sample size under different conditions (combinations of the three variables). The programming language that I use to perform my simulation is R (version 3.6.3) with Rstudio as my IDE (Integrated Development Environment). Below, I will only briefly introduce my program flow without every detail due to the space consideration.

What I actually simulated is a 10-year follow-up study which is going to estimate a mortality rate in the end. I manipulated the three variables to make several combinations which means the different conditions of the simulated studies. I have tried 100 conditions (population mortality rate in {0.01, 0.03, 0.05, 0.07, 0.09} x proportion of loss of follow up in {0.1, 0.2, 0.3, 0.4, 0.5} x level of association in {0.04, 0.06, 0.08, 0.10}). For each study under its particular condition, I further simulated its estimated mortality rate along different sample size (from 50 to 200). Given each particular condition and sample size, I performed the simulation for 1000 times and take the average of these 1000 simulated estimates as the point estimate made from this particular study.

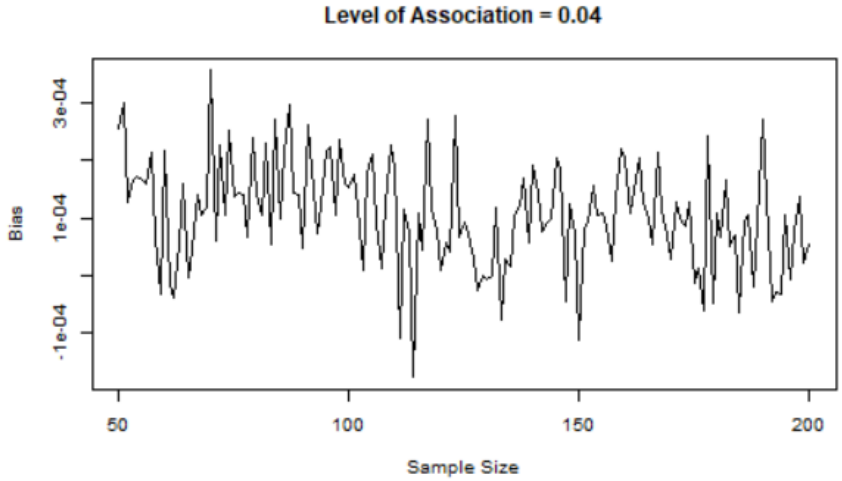
Through this way, I have generated several graphs of this kind:



The graph above illustrates an example of simulation for a study under a particular condition. The x-axis refers to the “sample size” and the y-axis refers to the “estimated mortality rate”. As for the population incidence rate (real incidence rate), it is marked by the horizontal line with the height of 0.01 in this example. The two dash lines then refer to the upper limit and lower limit of the 95% confidence interval for

each estimation along the sample size.

Furthermore, I have also generated several graphs to show the average of the bias each estimation produced along the sample size, which are of this kind:



The graph above then illustrates an example which shows the average bias produced by each estimation along the sample size under a particular condition. Each average bias is calculated by the formula:

$$\frac{\sum(Estimated\ incidence\ rate_i - Population\ incidence\ rate)}{Times\ of\ simulation}$$

The subscript of 'i' denotes the ith estimated incidence rate comes from the ith time of simulation, with the particular sample size.

In my simulation research, there are actually three assumptions for reviewers to pay attention. First, I assume that the population mortality rate is consistent under the whole study period, which is usually an assumption taken by a typical follow-up study. Second, I assume that the number of mortality in each year follows a poisson distribution with the parameter lamda equal to the product of the number of the remaining subjects and the population mortality rate. This assumption is taken since I assume "mortality" to be a rare event with low probability of occurrence. Third, I assume that the number of loss of follow-up in each year follows a binomial distribution with two parameters size and probability equal to the number of the remaining subjects and the proportion of loss of follow-up respectively. This assumption is taken since I assume "loss of follow up" not to be a rare event, which corresponds to the topic of this report—"follow-up study with loss of follow-up".

Result

First, please look at the appendix 1 for the several line charts which picture the main effect of the population mortality rate and the interaction between the population mortality rate and the level of association (given the proportion of loss of follow up as

0.1). There is a clear pattern that as the population mortality rate rises, the estimates become more vulnerable to underestimation, says that the level of association at which the underestimation occurs becomes lower. In fact, the overall degree of underestimation also seems to increase as the population mortality rate goes higher. In addition, it is obvious that once the underestimation occurs, the severity of it, say the degree of underestimation, increases along the sample size. The bar charts showed in the appendix 2 might provide a more comprehensive view of these patterns.

Then, please look at the appendix 3 for the several line charts which picture the main effect of the proportion of loss of follow-up and the interaction between the proportion of loss of follow-up and the population mortality rate (given the level of association as 0.1). There is an interesting pattern that as the proportion of loss of follow-up increases, the phenomenon of underestimation seems to diminish, which might violate one' intuition since one might expect that the underestimation should become more serious if one has relatively high proportion of loss of follow-up in one's study. In addition, the diminishment of underestimation along the sample size seems to become more clear as the population mortality rate rises (actually until the population mortality rate goes to around 0.1).

Discussion

What one might worry when performing estimation from a follow-up study with loss of follow-up is that the level of association might impair the independence of one's sampling, which might cause adverse effect on one's estimate. Indeed, as my simulation displays, the overall validity decline as the level of association increases, while the overall precision increases in a inverse fashion. This implies that one might come up with a relatively high probability to attain an "biased" estimate at least under some particular conditions, where one should especially take cautions.

To say that extra cautions should be taken even when the level of association is small, I manipulated the level of association with the maximum of 0.1, which means that there would be at most 10% of the mortality to be mistaken for the loss of follow-up in each year. Simulation under this kind of setting shows that underestimation might happen, implying that bias might be easy to occur whenever there is certain amount of loss of follow-up. The further point one should pay attention is that increasing one's sample size might not necessarily be a good idea to deal with the potential underestimation since the bias might go even greater though the precision goes higher, at least under some particular conditions.

To say that extra cautions should be taken even when the proportion of loss of follow-up is small, I digged into the interaction between the proportion of loss of follow-up and the sample size , and it seems to come up with a diminishing effect of sample size as the proportion of loss of follow-up increases. That is, a smaller

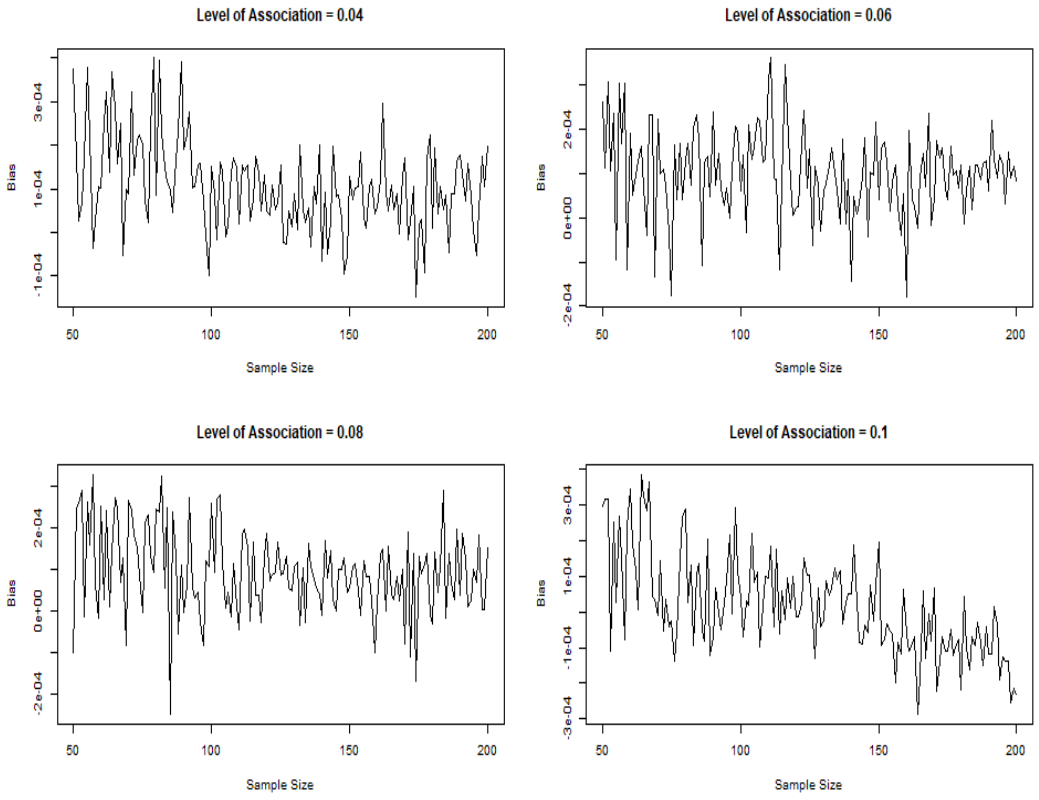
proportion of loss of follow-up might contribute more to the underestimation than a larger proportion of loss of follow-up, implying that one should still take cautions even though the proportion of loss of follow-up seems to be small. In addition, one should also pay attention that the diminishing effect of the sample size might become even more significant as the population parameter increases, at least under some particular conditions.

Conclusion

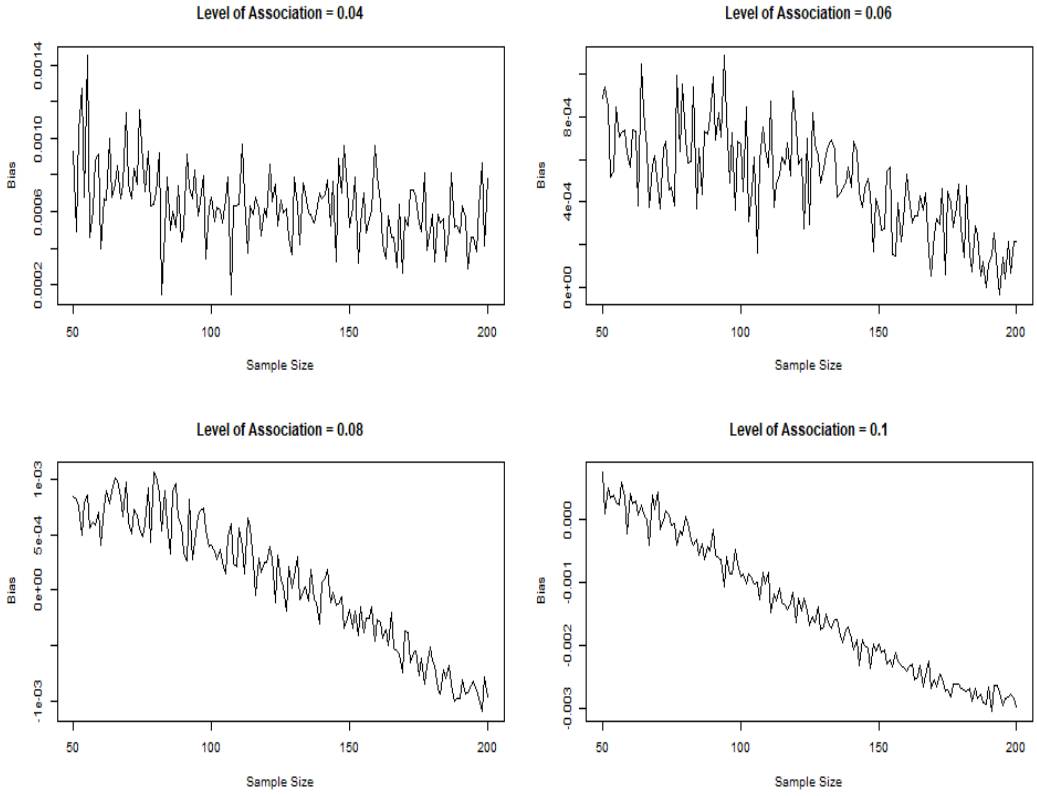
My simulation result is right an example, illustrating two import points that one should always keep in mind whenever he/she is going to perform estimation from a follow-up study with loss of follow-up. The first point is that large sample size may not always guarantee a good estimation especially under the condition where the population parameter and the level of association is small, even without a really high proportion of loss of follow-up. The second point is that one should always keep cautious even when the proportion of loss of follow-up is small, say < 20%, since the estimate made from a study with lower (< 20%) proportion of loss of follow-up might be in fact a more underestimated one than that made from a study with higher (> 20%) proportion of loss of follow-up, at least under the condition where population parameter and the level of association is small.

Appendix 1.

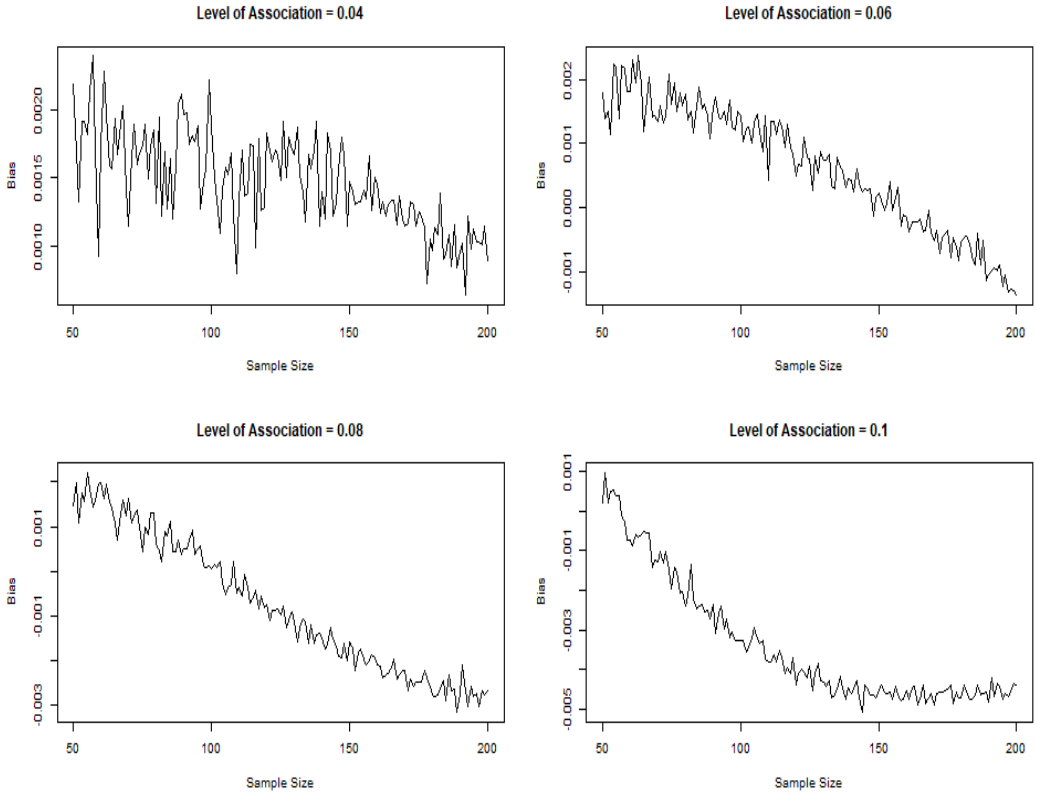
Population mortality rate = 0.01



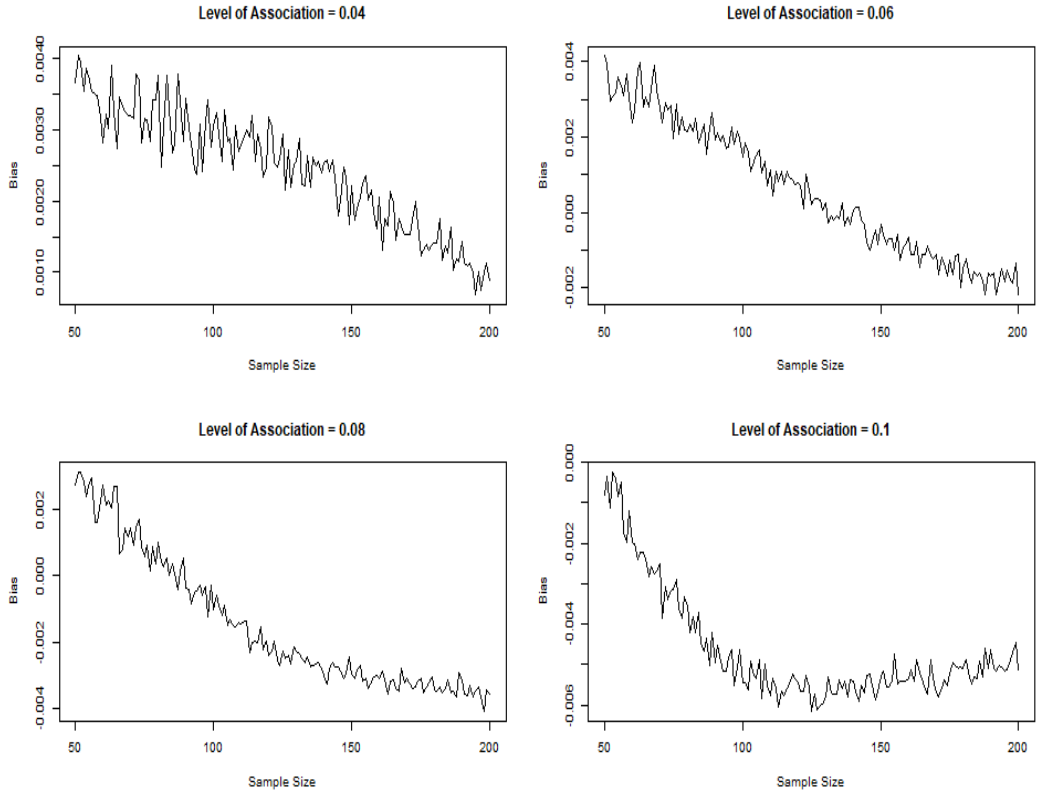
Population mortality rate = 0.03



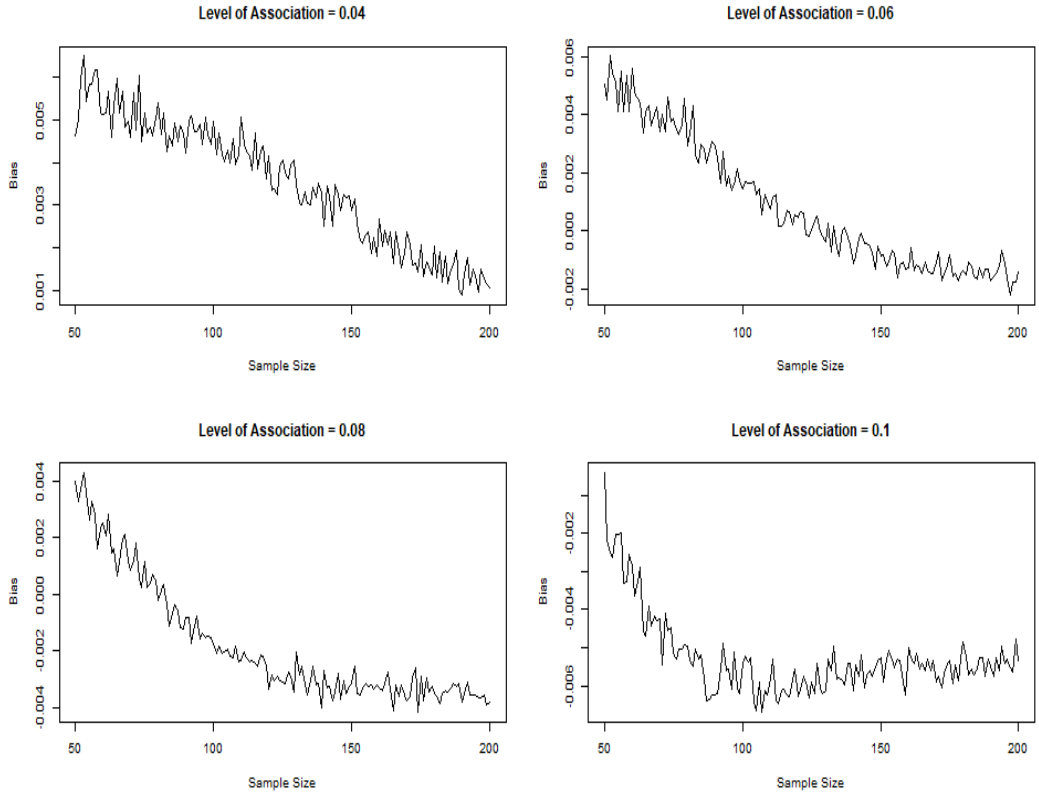
Population mortality rate = 0.05



Population mortality rate = 0.07

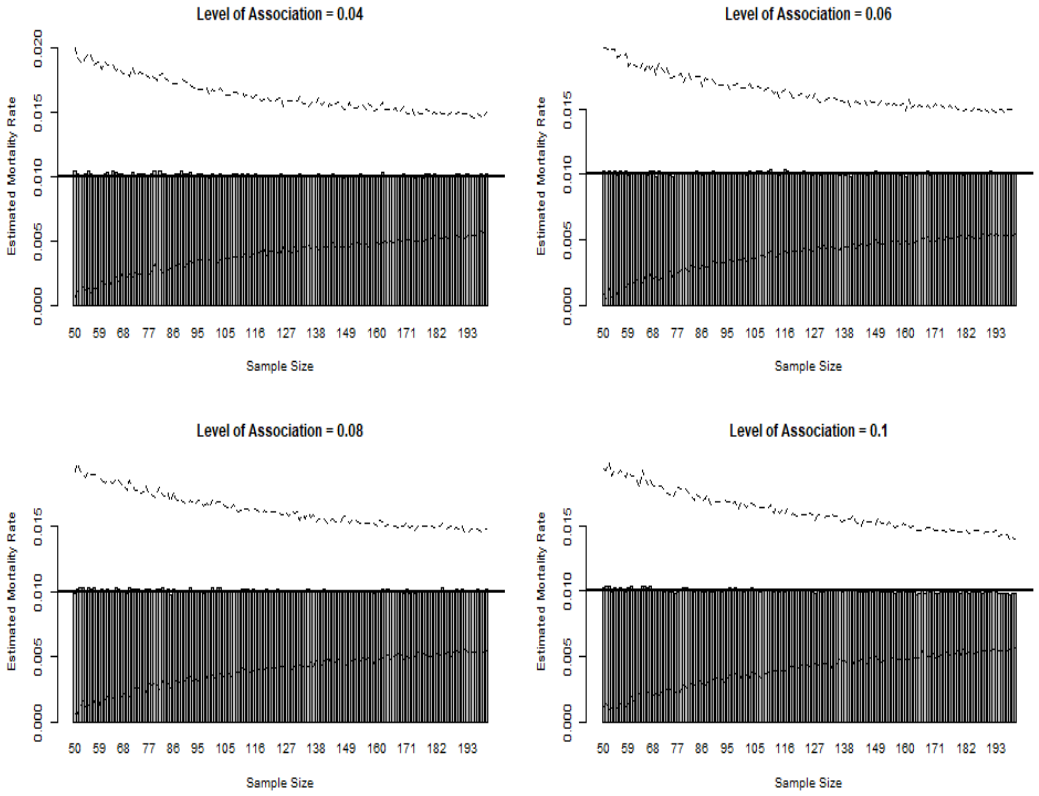


Population mortality rate = 0.09

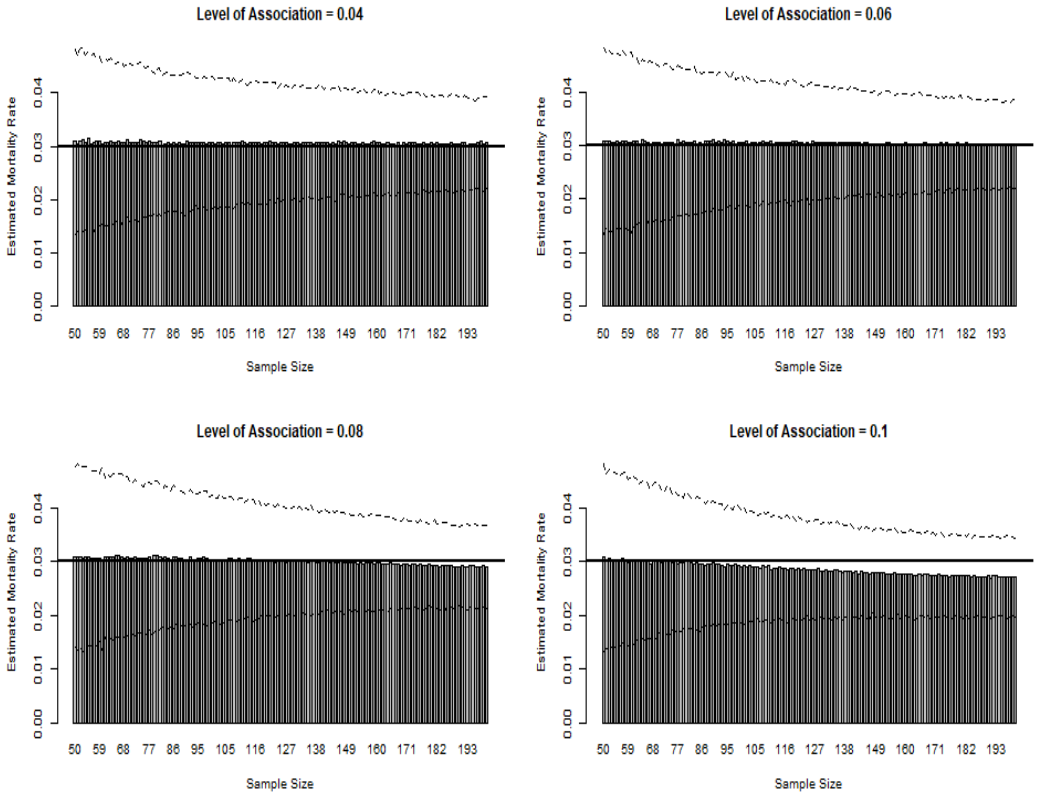


Appendix 2.

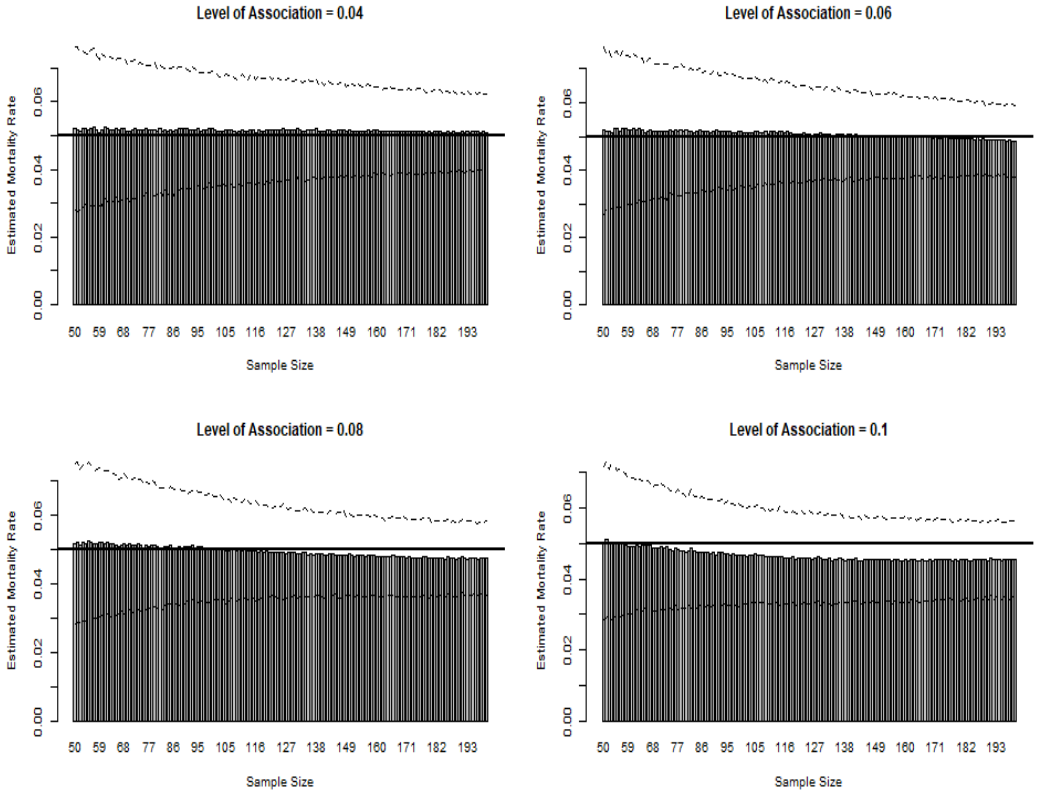
Population mortality rate = 0.01



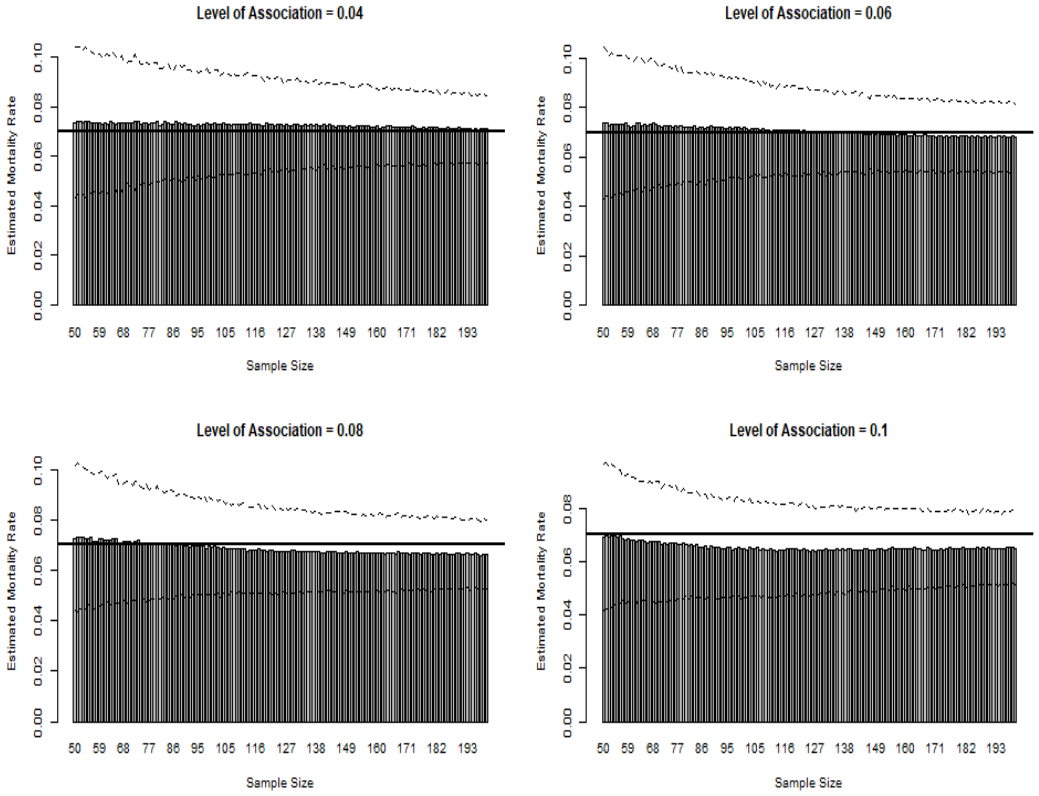
Population mortality rate = 0.03



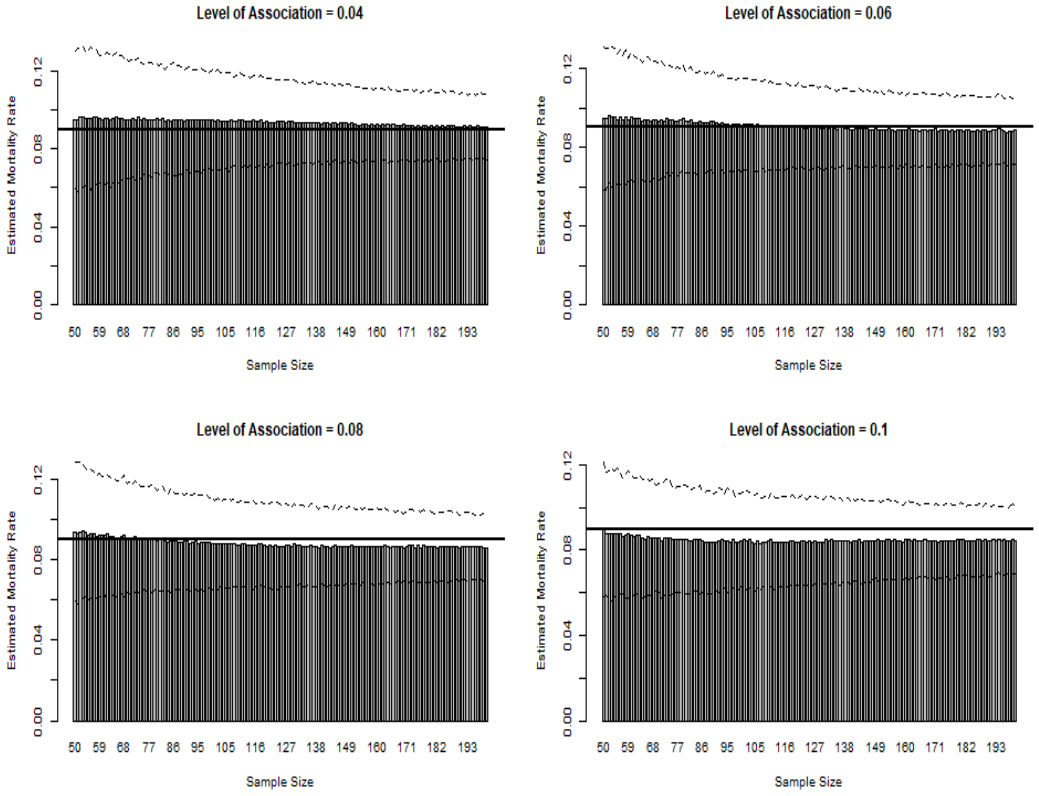
Population mortality rate = 0.05



Population mortality rate = 0.07

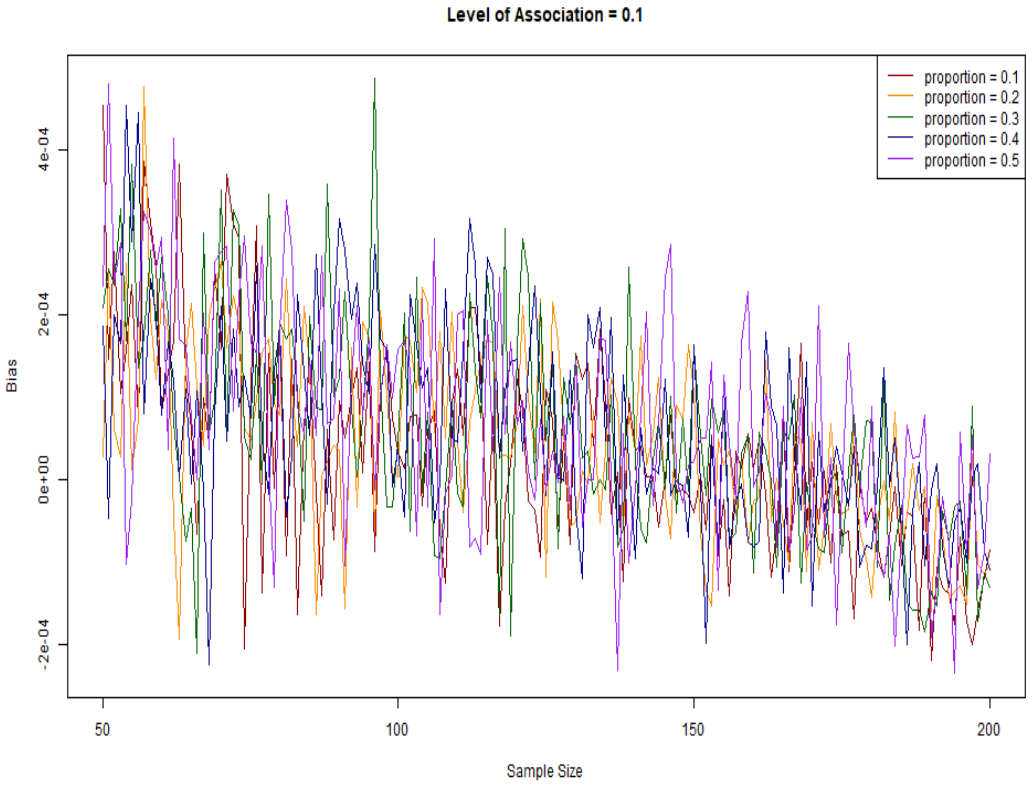


Population mortality rate = 0.09

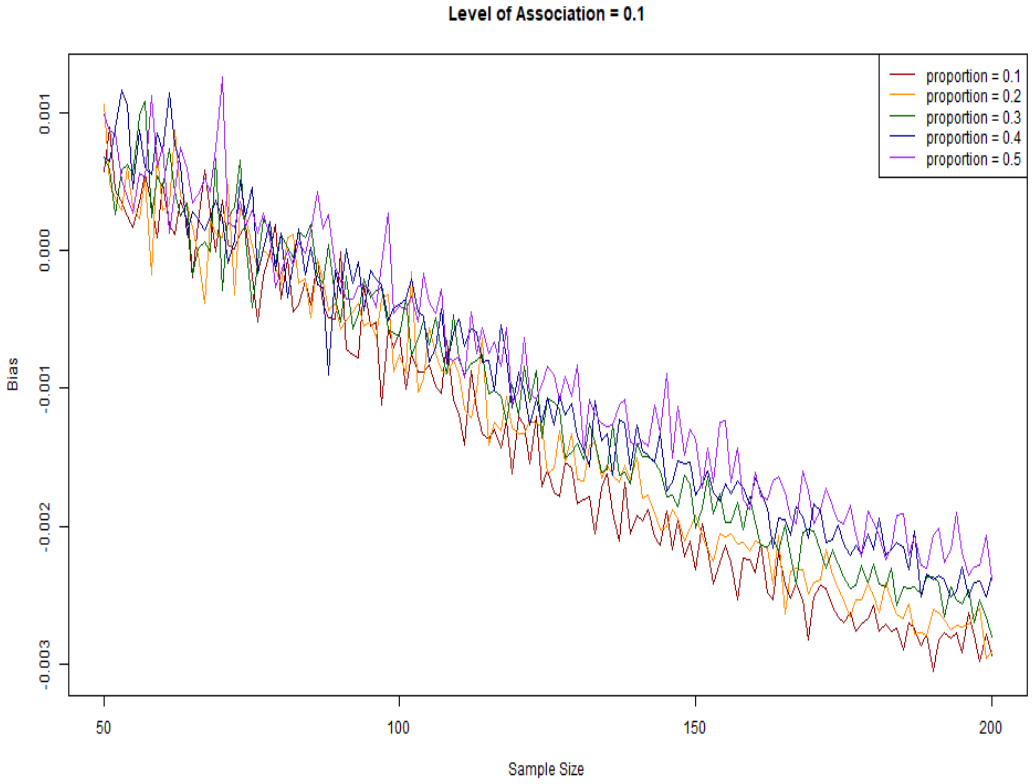


Appendix 3.

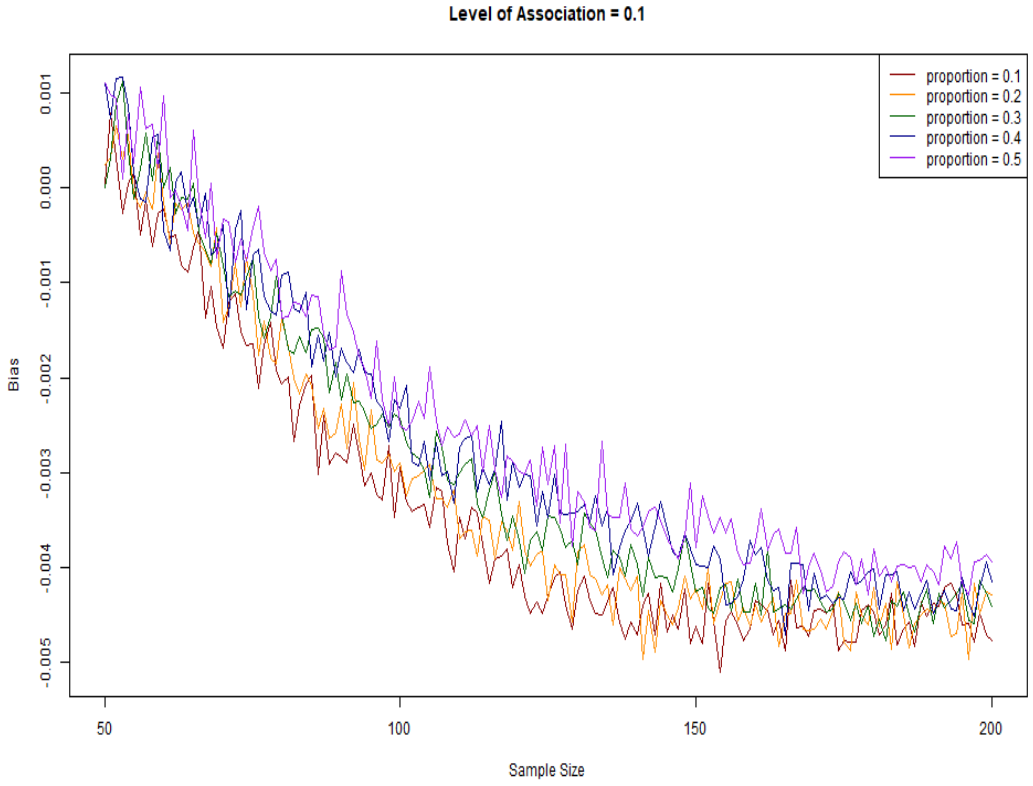
Population mortality rate = 0.01



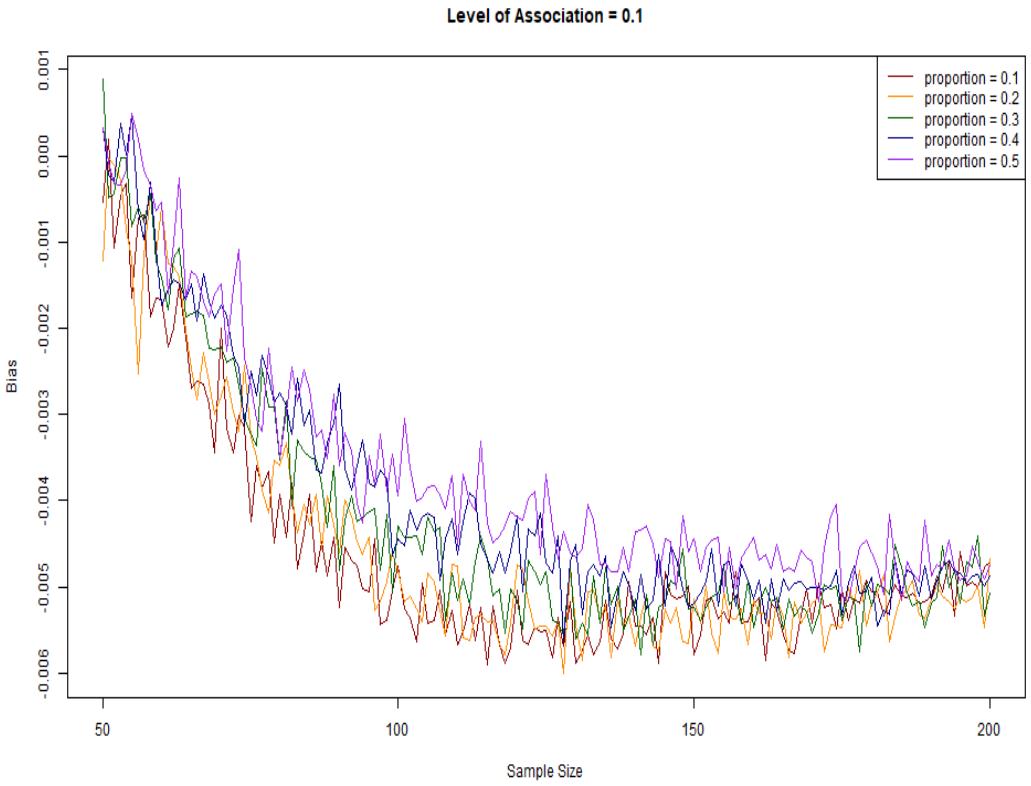
Population mortality rate = 0.03



Population mortality rate = 0.05



Population mortality rate = 0.07



Population mortality rate = 0.09