UPPSALA
UNIVERSITET

Cutoff Sample Size Estimation For Survival Data: A Simulation Study

By Huiwen Che

Department of Statistics

Uppsala University

Supervisor: Inger Persson, Katrin Kraus

Master Thesis 15 hp

2014

Abstract

This thesis demonstrates the possible cutoff sample size point that balances goodness of estimation and study expenditure by a practical cancer case. As it is crucial to determine the sample size in designing an experiment, researchers attempt to find the suitable sample size that achieves desired power and budget efficiency at the same time. The thesis shows how simulation can be used for sample size and precision calculations with survival data. The presentation concentrates on the simulation involved in carrying out the estimates and precision calculations. The Kaplan-Meier estimator and the Cox regression coefficient are chosen as point estimators, and the precision measurements focus on the mean square error and the standard error.

Keywords: sample size; simulation; survival analysis

Contents

Chapter 1

Introduction

1.1 Simulation and Sample size

Sample size estimation is an integral part of planning a statistical study. Usaully, a trade-off between accuracy of estimation and study cost is made in sample size decision. An adequate sample size is important to yield statistically significant results. A large sample size, however, may run over budget. Thus, a sample size that satisfies both aspects is required. Besides sample size, the follow up time, where careful thought is also given in medical research, is taken into consideration in this thesis. The longer the follow up time, the more information we know about the life expectancy. The right censoring, referring to the time of an obersvation's occurrence for the event greater than the specified study time, is often present in survival data due to the insufficient follow-up.

Statisticians calculate the required sample size based on the purpose of the study, the level of confidence, and the level of precision. The sample size analytic fomulas is one method to determine the sample size and the alternative way to estimate the sample size is simulation. Zhao and Li (2011) infered in their article that, the simulation technique, accommodating more complicated statistical designs, has increased use in sample size specification. The availability of computer simulation tools has driven the extensive use of computing intensive methods. The Monte Carlo simulation (referred as simulation in this thesis) can deal with uncertainty and it attempts to mimic the procedure samples collected from the population, which supports its use in sample size and parameter estimation (Efron and Tibshirani, 1986).

The motivation for this thesis is to demonstrate the use of simulation in precision and sample size estimations by example. By simulating on concrete data, I intend to illustrate practical results, confirming the theoretical analysis. For simplicity, the survival data used in this motivating example is regarded as the target population, where samples are drawn from. In this situation, the population parameter can be determined. With different sizes of samples taken from the population, statistical inferences are conducted, including sample statistics and precision estimation. The simulation plays a role as a procedure for evaluating the performance of different sample sizes. The multiple follow up time plans lead to multiple situations to repeat the analysis. The longer the follow-up, the greater number of events. It is easy to expect that as the sample size gets larger, the estimations will get more accurate and the same result applies to longer follow up time. However, it is of interest that of which cutoff sample size point or sample size region that we reach a certain desired precision, that we get dramatic performance improvement of the specified estimator, and that we do not have to sacrifice an increased sample size and thus increased cost to achieve a corresponding precision. It is possible to evaluate the effects of longer follow up time as well. The similar procedures are done in some pilot studies to estimate sample size required, but generally, these pilot studies are either more parametric oriented, assuming certain probability distribution to simulate from, or dependent on analogous previous studies. Teare et al. (2014), in their paper, compared the precision (the width of confidence interval) when sample sizes are different and suggested the recommended sample size in randomized controlled trials by sampling from distributions. In another paper, Lee et al. (2014) presented an real data example, whether the pilot 3 month data for 40 patients would proceed to main study of 233 patients at certain significant levels. The virtual scenario in my thesis may result in less adaptive but more detailed findings.

In this thesis, I outline the simulation method and report the results from the simulation study of statistics when applied to the Kaplan-Meier estimation and the Cox regression. In the final part, I make some conclusion remarks and a brief discussion. The simulation and analysis are implemented in software R.

1.2 Background

In this study, the population is generated based on the data that was reported by Kardaun (1983). In Kardaun's study, survival time of 90 males with laryngeal cancer who were diagnosed and treated during the period 1970-1978 was studied. Analogous to the original data of Kardaun's, a made-up population of 994 patients, including stage of cancer, year of diagnosis, month in which the patient was diagnosed, patient's age at diagnosis, and the survival time (measured in months), is used to conduct the simulation study. The male patients in this ficticious population (hereafter referred to as population) were diagnosed with laryngeal cancer during 1990-1998. There are four stages of cancer, among which stage 4 is of the highest severity and stage 1 is of the lowest severity.

All patients died within 9 years after the end of diagnosis year (i.e.within 9 years from 1998) in the population. To investigate follow up time effects on the study, I assume two study termination dates, Jan 2004, and the day when the last survivor died. Thus respectively, there are two populations with different study termination date. In population 1, the so-called censoring is detected; while in population 2, following all patients untill death occurs to every individual, complete survival time is recorded.

Censoring comes in the form of right censoring in this case, an observation terminated before the event occurs. In the laryngeal cancer data, if a patient's survival time (in months) T is greater than the study follow up time, namely, a survivor at the end of the study. This is marked as a censored observation. We do not obtain the survival time information after the follow-up. In the right-censored data, an observation's time on study is the time interval between diagnosis and either death or the end of the study, and the associated indicator of death (indicator of 1) or of survival (indicator of 0) are included.

The two populations have been set up. In population 1, the number of events is 915 (approximately 8$\%$ censoring); while in population 2, the number of events is 994 with no censoring. The two simulation pools will affect the information the samples inherit. Samples drawn from population 1 are anticipated to carry less information than samples from population 2.

Chapter 2

Method

To find out the possible threshold sample size value, we estimate the precision of inferences and specify a certain level of precision that is likely to achieve both estimation accuracy and cost efficiency. The measurements used to realize the evaluation are the mean squared error that incorporates the variability and bias of an estimator, and the standard error that is a typical measurement of precision. The mean squared error and the standard error are simple but useful performance measurements, which are also essentially associated with the simulation in the thesis. The particular simulation methodology used is the bootstrap method introduced by Efron (1979). The basic idea to estimate the standard error is that we generate a number of bootstrap samples of the same size by drawing randomly from the known observations, and calculate the estimator of interest for each bootstrap sample. When we derived a number of estimates of the estimator from bootstrap samples, we can estimate the standard error with respect to the estimator of interest. Efron and Tibshirani (1986) has shown that the boostrap estimate of standard error approaches to the true standard error as the simulation times are sufficient. The same idea can be adopted when calculating the mean squared error.

In terms of the estimators, the thesis considers models for survival analysis which have the following three main characteristics: (1) time-to-event data features; (2) censored observations; (3) the effect of explanatory variables on the death time. The Kaplan-Meier estimator, accommodating characteristics (1) and (2), and the Cox regression, accommodating characteristics (2) and (3), are selected to analyse the survival data. The ability of Kaplan-Meier method to summarize survival probability intuitively when there is censoring and to offer further implications in survival analyses is the prominent reason to devote effort to evaluate the Kaplan-Meier estimator. While the Kaplan-Meier method focuses more on the basic shape of survival function, the Cox regression proceeds to further complicated analysis of the relationship between survival time and explanatory variables. Since the Cox regression is also the main model used in Kardaun's study for analysis and the made-up population is ground on the data from Kardaun's study, the thesis continues to use the Cox regression to model the survival data. The survival models used are based on the following basic definitions.

2.1 Survival function and Hazard function

In survival analysis, it is common to employ survival function to describe the time-to-event phenomena (Klein and Moeschberger, 1997). The survival function is defined as

$$S(x) = Pr(X > x). \tag{2.1}$$

If the event of interest is death, the survival function models the probability of an individual surviving beyond time x. $S(x)$ is bounded between 0 and 1 as a probability. A closely related function is the cumulative distribution function of a random variable X, which is defined as the probability that X will be less than or equal to x.

$$F(x) = P(X \leq x). \tag{2.2}$$

If X is a continuous random variable, $S(x) = 1 - F(x)$, and $S(x)$ is non-increasing monotone function.

For continuous survival data, we want to quantify the risk for event occurrence at exactly time t (Klein and Moeschberger, 1997), and hence the hazard function is defined by

$$h(x) = \lim_{\Delta x \to 0} \frac{P[x \leq X < x + \Delta x | X \geq x]}{\Delta x}. \tag{2.3}$$

The hazard function, or simply hazard rate, is nonnegative. It can be written as

$$h(x) = -\frac{d}{dt} \log S(x). \tag{2.4}$$

2.2 Kaplan-Meier method

The Kaplan-Meier (KM) estimator, also known as the product-limit estimator, is widely used in estimating survivor functions. Kaplan and Meier (1958) gave a theoretical justification to the method by showing that the KM estimator is a nonparametric maximum likelihood estimator. The estimator is defined as:

$$\hat{S}(t) = \begin{cases} 1 & \text{if} \quad t < t_1, \\ \prod_{t_i \leq t}[1 - \frac{d_i}{Y_i}] & \text{if} \quad t_1 < t. \end{cases} \tag{2.5}$$

where t_1 is the first observed failure time, d_i is the number of individuals who died at time t, and Y_i is the number of individuals who are at risk of the event of interest. The KM estimator also takes censoring into account. When there is censoring, being at risk means that individuals have not experienced the event nor have they been censored prior to time t_i. Thus Y_i is the number of survivors substracting the number of censored observations.

Figure 2.1 and 2.2 shows the graph of the Kaplan-Meier estimates of survival function for population 1 and population 2 respectively.

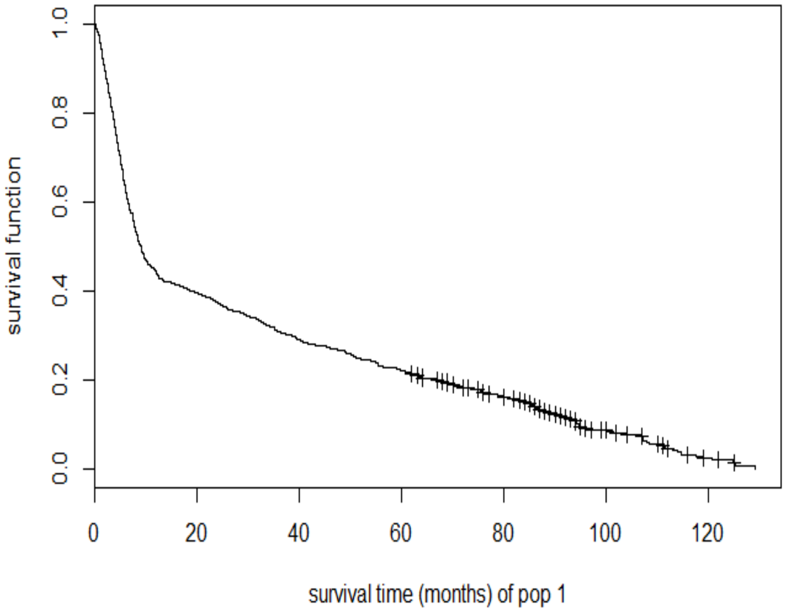

Figure 2.1: Kaplan-Meier survival function for population 1. The small verticle tick-marks indicate censoring

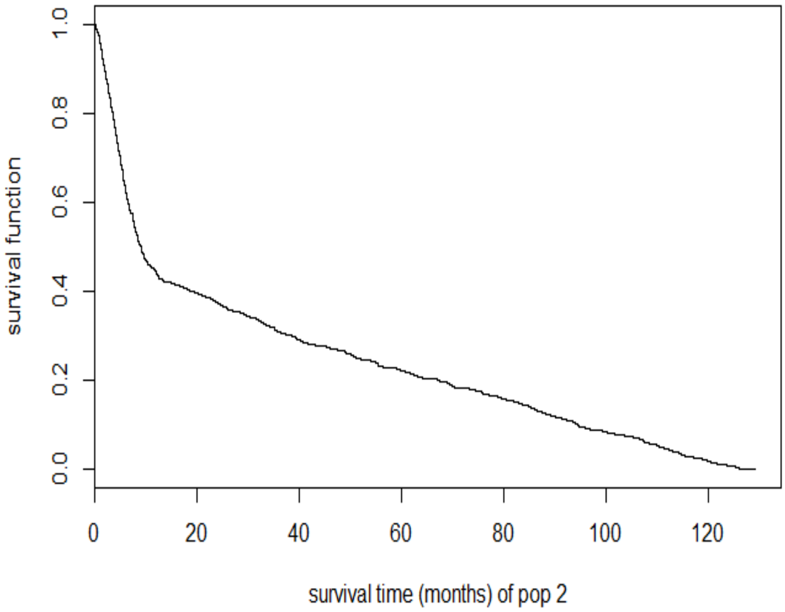

Figure 2.2: Kaplan-Meier survival function for population 2

The Kaplan-Meier method produces estimates of survival function at the various death times. In this thesis, special interest is given to the survival probabilities associated survival time (months) - the quartile estimates (point estimate on survival probability of 75%, 50%, and 25%). The summary parameters of the two populations for time is illustrated in Table 2.1. As the survival time gets greater, the survival probability declines. The survival times of the two populations at each of the three probability points are in register. Sample estimates will be obtained to compare with the following true parameters.

Survival probability	75%	50%	25%
population 1	4.20	8.95	51.20
population 2	4.20	8.95	51.20

Table 2.1: Quartile estimates of survival times (months)

2.3 Mean squared error

The mean squared error (MSE) measures the mean squared difference between the estimator and the parameter and it evaluates the error made by the estimator, which serves as a measurement of goodness of an estimator (Casella and Berger, 2002, chap. 7). The MSE has the interpretation

$$MSE(\hat{t}) = \mathrm{E}(\hat{t} - t)^2 \tag{2.6}$$

the MSE can be decomposed into a sum of bias and variance

$$MSE(\hat{t}) = \mathrm{E}(\hat{t} - t)^2 = Var(\hat{t}) + B(\hat{t})^2 \tag{2.7}$$

the variance measures the precision of the estimator, while the bias - the difference between the true survival time and the mean of estimated survival times, measures the accuracy of the estimator.

2.4 Cox regression

Survival analysis is typically concerned with examining the relationship of the survival distribution to some covariates. Cox regression modelling is a modelling approach to explore the effects of variables (so-called covariates) on survival, as Fox (2002) described in his article. The prediction idea in survival regression is similar to that in ordinary regression (Klein and Moeschberger, 1997). The non-parametric strategy that leaves the baseline hazard $h_0(t)$ unspecified is used here to regress the survival times on the explanatory variables. The model, also called proportional hazards model, was proposed by Cox (1972) as follows

$$h_i(t) = h_0(t)exp(\beta_1 * x_{i1} + \beta_2 * x_{i2} + \cdots + \beta_k * x_{ik}) \tag{2.8}$$

where $h_0(t)$ is an arbitrary baseline hazard rate; that is when all covariates are set to zero at time t. $X_i = (x_{i1}, \cdots, x_{ik})$ are the covariates (risk factors) for the ith individual, and $\beta = (\beta_1, \cdots, \beta_k)$ are regression coefficients that predict the proportional change in the hazard. The covariates $(\beta_1 * x_{i1} + \beta_2 * x_{i2} + \cdots + \beta_k * x_{ik})$ form the model linearly. Suppose two individuals i and i', the associated linear parts are as follow

$$\eta_i = \beta_1 * x_{i1} + \beta_2 * x_{i2} + \cdots + \beta_k * x_{ik}$$

and

$$\eta_i = \beta_1 * x_{i'1} + \beta_2 * x_{i'2} + \cdots + \beta_k * x_{i'k}$$

The hazard rates in the Cox model are proportional, as the quantity 2.7 demonstrates.

$$\frac{h_i(t)}{h_{i'}(t)} = \frac{h_0(t)exp^{\eta_i}}{h_0(t)exp^{\eta_{i'}}} = exp(\eta_i - \eta_{i'}) \tag{2.9}$$

which is a constant. An individual with risk factor X_i experiencing the event as compared to an individual with risk factor $X_{i'}$ is $exp(\eta_i - \eta_{i'})$.

The explanatory variables of interest in this thesis are stage, age, and year of diagnosis. Using the Cox model, the hazard at time t is expressed as

$$h_i(t) = h_0(t)exp(\beta_1 * stage + \beta_2 * age + \beta_3 * yearofdiagnosis)$$

or, equivalently,

$$\log h_i(t) = \log h_0(t) + \beta_1 * stage + \beta_2 * age + \beta_3 * yearofdiagnosis$$

Tables 2.2 and 2.3 show the parameter estimates of the Cox regression after fitting the model to population 1 and population 2. All three covariates have statistically significant coefficients. The regression coefficients of the two populations have nuances and the standard errors of the coefficients for population 2 are slightly smaller than those for population 1 due to the censoring in population 1. The exponentiated coefficients represent the multiplicative effects on the hazard. For instance, as shown in Table 2.2, with an additional stage of the cancer and other covariates held constant, the hazard (risk of dying at the next instant) increases by a factor of 1.344 or 34.4 percent. Holding other covariates constant, an additional year of diagnosis reduces the hazard by a factor of 0.935 or 6.5 percent.

	coef[1]	exp(coef)[2]	se(coef)[3]	z[4]	p[5]
stage	0.2959	1.344	0.03296	8.98	0.0e+00
age	0.0166	1.017	0.00331	5.02	5.3e-07
year[6]	-0.0668	0.935	0.01556	-4.30	1.7e-05

Table 2.2: The Cox regression on population 1

[1] coefficient

[2] exponentiated coefficient

The coefficients for population 2 are similar to those for population 1. However, the presence of 8% censoring in population 1 causes some differences. Viewing at the exponential coefficients from two tables (Table 2.2 and 2.3), we find that the hazards of population 1 are of greater increase or decrease than the hazards of population 2. Taking the covariate - stage as the example, the hazard of population 1 increases 34.4 percent, while the hazard of population 2 increases 33.6 percent with an additional stage of the cancer and holding other covariates constant. The observation is also true for the covariate - age, though the difference is tiny. For the covariate - year of diagnosis, the hazard of population 1 (6.5 percent) reduces more than the hazard of population 2 (4.7 percent), which interpreting in another way, we may conclude that the impact of year of diagnosis is inflated. It seems that the risk of dying is overestimated in population 1 due to censoring, comparing with population 2.

	coef	exp(coef)	se(coef)	z	p
stage	0.2895	1.336	0.0317	9.13	0.0e+00
age	0.0159	1.016	0.0032	4.98	6.4e-07
year	-0.0478	0.953	0.0147	-3.25	1.2e-05

Table 2.3: The Cox regression on population 2

[3] standard error of coefficient

[4] Z-score

[5] P-value

[6] year of diagnosis

Chapter 3

Simulations: Kaplan-Meier Estimation

In the following two chapters, simulation procedures and results are discussed. This chapter presents estimates of survival function, with different sample sizes drawn from the population. The nonparametric Kaplan-Meier estimator is used here. As stated in the previous chapter, quartiles of KM estimates when the survival probabilities are 0.75, 0.50, and 0.25 are the primary consideration for each simulation.

3.1 Simulation design

The simulation steps are shown below

1. Generate random index of size n with replacement.

2. Draw a sample X of n observations from population 1, according to the index generated in step 1.

3. Calculate KM estimators from the drawn sample, extract the quartiles estimates, and store these three estimates.

4. Repeat steps 1 to 3 for 5000 times.

5. Each simulation has 5000 estimated survival time at each survival probability and calculate the mean , variance, and bias of the estimates at each quartiles.

6. Repeat steps 1 through 5 for a range of different sizes (n = 30, 40, 50, 60, 75, 100, 125, 150, 200, 250, 300, 350, 400, 450, 500).

7. Repeat the above procedures using population 2.

In each simulation, the associated time t_1, t_2, t_3 to $S(t_1) = 0.75, S(t_2) = 0.50, S(t_3) = 0.25$ are stored. To capture the average performance of the estimator, we consider the MSE. The variance of survival times at each targeting survival probability point, the average bias and the MSE are computed in the simulation.

3.2 Results

Implementing the above procedures, the resulting estimates appear in Tables 3.1 and 3.2. First, take a close look at the results from population 1 shown in Table 3.1. When comparing the bias, horizontally (at the three probability points when sample size is the same), the high survival probability point is more likely to have lower bias, although there are a few exceptions at the 50% survival point; and vertically (at different sample size points), the main trend is that the greater the sample size, the smaller the bias. The difference between bias, however, is quite insignificant. In this simulation study, the main source of bias seems to arise from the non-representative sample. With greater sample size, the sample could be more representative. But since all biases are small, we may assume that the simulation setting actually plays a role in sampling representative samples. The remarkable difference lies in variance. Vertically, larger sample sizes indicate lower variance, which is most notable in the 25th percentile point. Horizontally, the variance soars up as survival time gets longer. One possible explanation for this phenomenon could be censoring. In population 1, the censored observations are gathered after survival time of 60 months, which corresponds to survival probability below the 25 percent (as shown in Figure 2.1 in the previous chapter). In the 25th percentile survival probability point, the less information about death is known, leading to less accurate estimation of survival time. Another reason may be that there is smaller number of observations at longer survival time. As patients die or get censored, less and less information is available, which leads to a larger variance.

The results derived from the population 2 in Table 3.2 share consistent trend with results from the population 1. The changing pattern of the performance in bias, variance and MSE is similar to Table 3.1. Comparing the two tables, it is hard to conclude any major difference due to the degree of censoring. There is probably one noteworthy exception, however, and that is the MSE or the variance at the 25 percent probability point. The differences of the variances

sample size	75%Bias	75%Var	75%MSE	50%Bias	50%Var	50%MSE	25%Bias	25%Var	25%MSE
30	0.11520	1.50823	1.52150	2.70359	51.30076	58.61016	0.68037	420.45954	420.92244
40	0.06980	0.98183	0.98670	1.98093	34.36520	38.28928	0.86410	303.80036	304.54702
50	0.07712	0.82087	0.82682	1.53499	23.73817	26.09436	0.81218	264.47523	265.13486
60	0.04677	0.65724	0.65942	1.28006	20.20743	21.84599	0.55657	218.89704	219.20681
75	-0.01316	0.52448	0.52465	0.96490	13.05787	13.98890	1.55563	187.39595	189.81593
100	0.01390	0.37531	0.37551	0.59644	7.21642	7.57216	0.47290	138.98899	139.21262
125	0.05390	0.30673	0.30963	0.50464	5.17277	5.42743	0.16310	112.62423	112.65083
150	0.01550	0.25358	0.25382	0.40566	3.99660	4.16116	0.77368	95.94227	96.54085
200	0.00210	0.18213	0.18214	0.20731	2.11572	2.15869	0.38242	70.95333	71.09958
250	0.00928	0.15339	0.15347	0.17881	1.63492	1.66690	0.52260	60.41012	60.68323
300	-0.00458	0.11771	0.11773	0.09620	1.14603	1.15529	0.50525	47.38854	47.64382
350	0.00254	0.10306	0.10307	0.10310	0.91237	0.92300	0.64590	41.09315	41.51034
400	-0.00203	0.08889	0.08889	0.09154	0.79341	0.80179	0.65155	35.53592	35.96043
450	-0.00364	0.07895	0.07896	0.03990	0.64337	0.64496	0.47350	32.13124	32.35544
500	-0.00138	0.07009	0.07009	0.03721	0.54959	0.55097	0.42431	27.69627	27.87631

Table 3.1: The quartiles KM estimates of bias, variance and MSE for population 1

between the two populations are relatively large, especially for some small sample sizes, such as sample size of 30. The variances at the 25 percent point of population 1 are greater than the variances of population 2, which may indicate less information in population 1 because of the censoring.

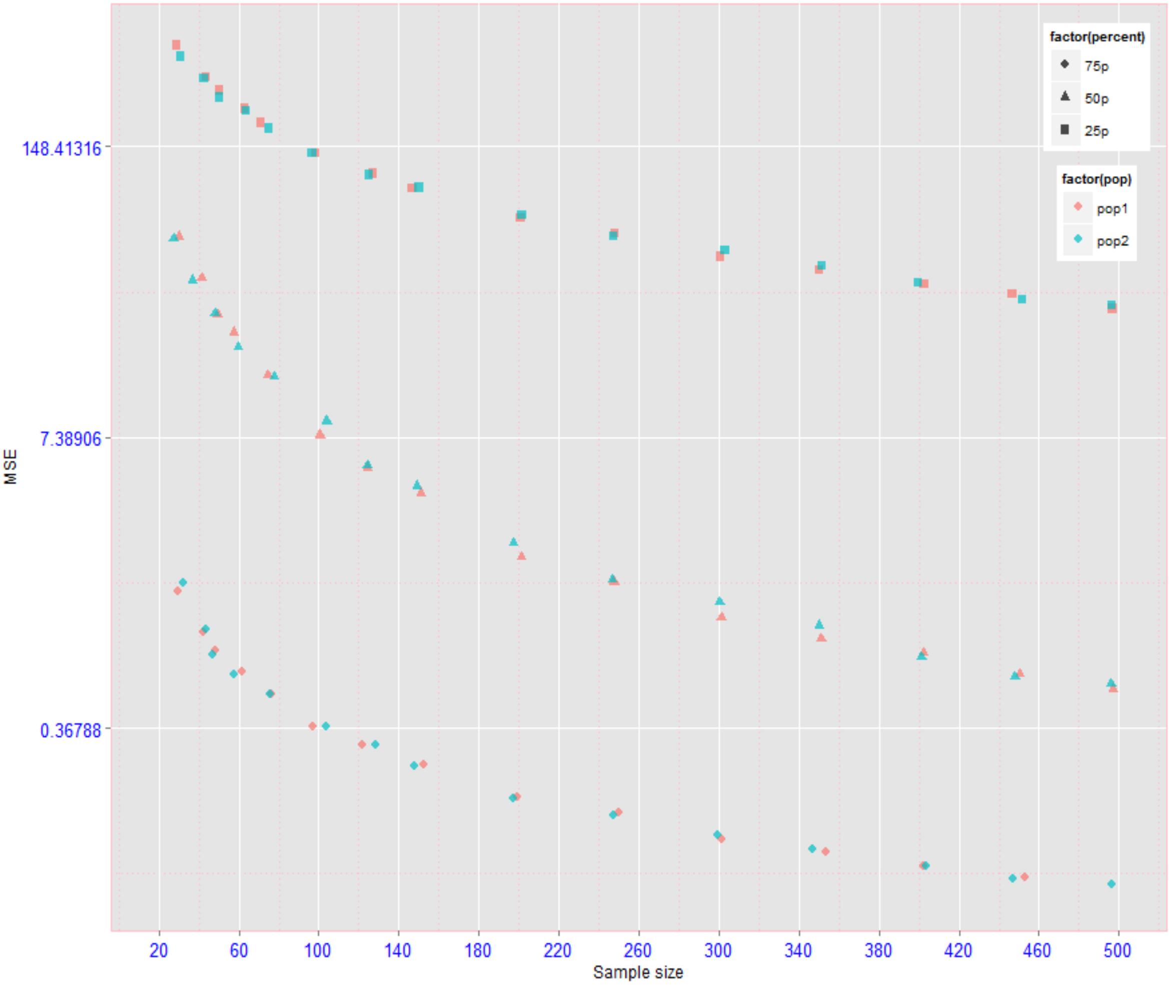

Figure 3.1: MSE of survival time at the three quartiles probability points for the two population. The vertical axis is on a logarithmic scale. The three probability points are in different symbols and the two populations are in different colors.

Figure 3.1 shows the MSE result graphically. The bias of the estimates are quite close (Table 3.1, 3.2) and the difference of MSE lies mainly in variance. Again, we see that the MSE is of greater value at the lower survival probability point and the differences between two populations are quite small in this logarithmic scaled graph[1]. The MSE declines as the sample

[1] The vertical axis is on a logarithmic scale due to the relative large range of the MSE values. The logarithms of the MSE values produce a more decent graph to see the trend.

sample size	75%Bias	75%Var	75%MSE	50%Bias	50%Var	50%MSE	25%Bias	25%Var	25%MSE
30	0.12456	1.62484	1.64036	2.56653	50.89364	57.48072	0.01022	375.54986	375.54997
40	0.06476	1.01728	1.02147	1.94487	33.60324	37.38576	0.14262	298.91219	298.93253
50	0.05998	0.77794	0.78154	1.54496	24.23997	26.62687	0.60056	245.34415	245.70482
60	0.04474	0.63842	0.64042	1.18701	17.30307	18.71206	0.23290	214.34067	214.39491
75	-0.01088	0.52251	0.52263	0.93484	12.92215	13.79608	1.61344	176.30919	178.91238
100	0.01791	0.37394	0.37426	0.67663	8.25865	8.71648	0.26768	138.95243	139.02408
125	0.03812	0.30754	0.30899	0.47986	5.29591	5.52617	0.00652	110.72287	110.72291
150	0.01450	0.25029	0.25050	0.38719	4.33859	4.48851	0.70214	96.88729	97.38030
200	0.00748	0.17726	0.17731	0.26129	2.42025	2.48852	0.70176	73.01640	73.50887
250	-0.00048	0.14877	0.14877	0.17477	1.67099	1.70154	0.60050	58.31568	58.67628
300	0.01455	0.12133	0.12154	0.14760	1.32389	1.34568	0.60266	50.67014	51.03334
350	-0.00358	0.10578	0.10580	0.10324	1.04873	1.05939	0.56174	43.19440	43.50995
400	-0.00236	0.08845	0.08846	0.05087	0.76262	0.76521	0.47388	36.29475	36.51931
450	-0.00562	0.07824	0.07827	0.03107	0.62564	0.62661	0.49254	30.30830	30.55090
500	-0.00293	0.07348	0.07349	0.05530	0.57760	0.58066	0.55070	28.58522	28.88850

Table 3.2: The quartiles KM estimates of bias, variance and MSE for population 2

size goes up, and meanwhile it is apparent that the declining rate of MSE slows down as the size grows. Though, there is no definition of best estimator in terms of the MSE, the sample size of 100 appears to be the threshold point of trade-off concerning all three probability points of interest. When the sample size is smaller than 100, the MSE curves have steeper slope. When the sample size is beyond 100, all curves are fairly flat. And the phenomenon applies to the both population estimations.

Chapter 4

Simulation: Cox Regression Estimation

4.1 Simulation design

The underlying distribution of the data is unknown, but the population is available. Therefore the sampling strategy used in this research is case sampling, sampling individual cases - each row of the data frame, to draw random samples. The simulation steps are as follow

1. Generate random index of size n with replacement.
2. Draw a sample X of n observations from population 1, according to the index generated in step 1.
3. Regress the sample X on the Cox regression model specified before and store the estimate of each coefficient - stage, age, and year of diagnosis.
4. Repeat steps 1 to 3 for 5000 times.
5. After the replication, there are 5000 estimates for each coefficient, calculate mean and standard deviation among these 5000 estimates, get the estimated mean and standard error of each coefficient.
6. Repeat steps 1 through 5 for a range of different sizes (n = 15, 17, 19, 22, 25, 28, 32, 36, 40, 45, 50, 55, 65, 75, 85, 100, 125, 150, 200, 250, 300, 350, 400, 450, 500).
7. Apply the above procedures to population 2.

As the population size is finite, if sample without replacement, the covariance of the different sample values is non-zero. To rule out the dependence, I sample with replacement. Sample sizes are selected from small sizes as 15 to relatively large sizes as 500. In this study, if the sample size is lower than 15, the number of events may be not enough to do regression. The perfomance of small sample sizes may change remarkably, and hence the intervals are small between chosen small sample sizes.

4.2 Results

For the design described in the previous section, the estimation of mean regression coefficients and their standard errors of varying sample sizes are derived from simulation. The selected outcome (sample size n = 15, 20, 30, 40, 50, 75, 100, 125, 150, 200, 300, 400, 500) of population 1, follow up time terminating on Jan 2004, is shown in Table 4.1.

sample size	stage	se.stage	age	se.age	year of diag	se.year of diag
15	0.4373869	0.4873281	0.0196584	0.0468380	-0.1003996	0.2197601
20	0.3862809	0.3451874	0.0189307	0.0350348	-0.0874264	0.1693540
30	0.3564377	0.2447388	0.0181800	0.0257017	-0.0838596	0.1229153
40	0.3439469	0.2047557	0.0178105	0.0208841	-0.0785292	0.0992787
50	0.3327355	0.1768086	0.0173763	0.0183326	-0.0771482	0.0864128
75	0.3167494	0.1420939	0.0172394	0.0145324	-0.0728275	0.0671305
100	0.3136281	0.1148034	0.0170618	0.0119367	-0.0712902	0.0562382
125	0.3088530	0.1031355	0.0169817	0.0104599	-0.0712314	0.0496486
150	0.3071950	0.0916850	0.0171219	0.0096201	-0.0682126	0.0453214
200	0.3031681	0.0770696	0.0172510	0.0082965	-0.0686442	0.0388819
300	0.3013235	0.0635650	0.0168542	0.0066123	-0.0686322	0.0311307
400	0.2990278	0.0554107	0.0167197	0.0056633	-0.0675111	0.0270559
500	0.2989356	0.0497359	0.0169062	0.0051131	-0.0666488	0.0239313
parameter	0.2959	0.03296	0.0166	0.00331	-0.0668	0.01556

Table 4.1: The estimated regression coefficients and standard errors for population 1 with study ending on Jan 2004

The estimated mean coefficients, on average, get closer to the population parameter (see Table 2.2) as the sample size increases. In regression modelling aspect, the hazards are overestimated when sample sizes get smaller, since the samller the sample size, the greater impact of explanatory variables. The performance improves dramatically among small sample sizes; while the results change slowly among large sample sizes. The standard error declines when the sample size increases as expected and it approaches to the population standard error. The sample size influences standard error considerablly for small sample sizes.

The similar results hold true for population 2 (see Table 2.3) with longer follow up time as shown in Table 4.2.

sample size	stage	se.stage	age	se.age	year of diag	se.year of diag
15	0.4261169	0.4958081	0.0210507	0.0458473	-0.0939805	0.2163588
20	0.3722668	0.3313938	0.0187941	0.0337496	-0.0804861	0.1600106
30	0.3484987	0.2436342	0.0175071	0.0249944	-0.0723029	0.1182873
40	0.3317312	0.1984947	0.0172376	0.0206801	-0.0658672	0.0975629
50	0.3241549	0.1738218	0.0167006	0.0174436	-0.0617790	0.0834075
75	0.3130634	0.1334420	0.0167675	0.0139466	-0.0581667	0.0646893
100	0.3057838	0.1121780	0.0164154	0.0116479	-0.0560753	0.0560396
125	0.3049019	0.0986926	0.0162281	0.0102005	-0.0555682	0.0477849
150	0.2981441	0.0890423	0.0163201	0.0091576	-0.0537220	0.0435515
200	0.2968767	0.0776443	0.0160546	0.0079510	-0.0525709	0.0369441
300	0.2947015	0.0609716	0.0161918	0.0063128	-0.0516554	0.0302071
400	0.2946066	0.0524316	0.0161907	0.0054369	-0.0505546	0.0261691
500	0.2924009	0.0459512	0.0159905	0.0049064	-0.0501157	0.0228183
parameter	0.2895	0.0317	0.0159	0.0032	-0.0478	0.0147

Table 4.2: The estimated regression coefficients and standard errors for population 2 with study ending on last death of cancer patients

Considering the two follow up time plans, Figures 4.1, 4.2, and 4.3 illustrate that as the sample increases, estimates associated with the two follow-up plans approach their own population parameters respectively. The censoring in population 1, which means fewer numbers of death, results in higher values of the parameters in population 1 than those in population 2. This further leads to the systematic estimation difference between samples drawn from two populations. The scale of the difference is fairly small, and tends to be neglectable. The standard errors between the two time plans, as shown in Tables 4.1 and 4.2, possess slight difference. Generally, the estimated standard errors in shorter follow-up are a little bit greater, comparing to the standard errors in longer follow-up, which is caused by limited death information in shorter follow-up. The neglectable difference in this case might owe to the low degree of censoring (approximately 8% censoring) difference between the two follow up time plans. We may expect some considerable effects of longer follow up time and of higher degree of censoring difference in other cancer studies.

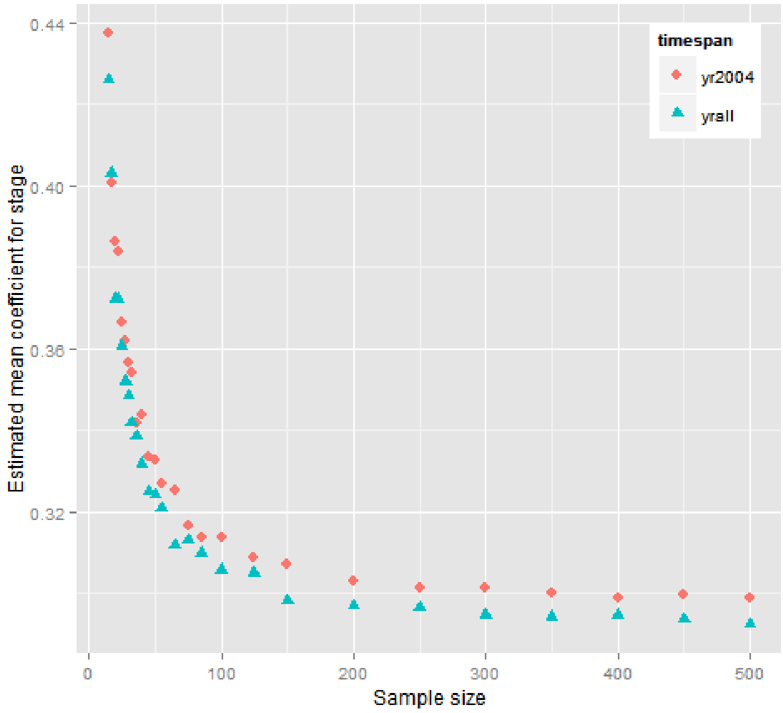

Figure 4.1: Estimated mean regression coefficient for stage. The points in red are estimates for population 1 (population parameter $\beta_{stage} = 0.2959$) and points in blue are estimates for population 2 (population parameter $\beta_{stage} = 0.2895$).

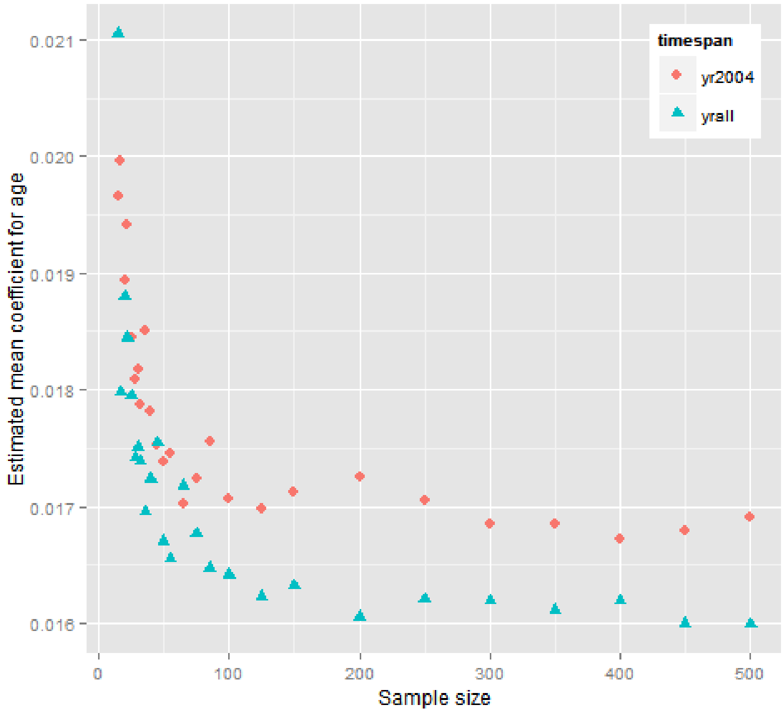

Figure 4.2: Estimated mean regression coefficient for age (population 1 parameter $\beta_{age} = 0.0166$ and population 2 parameter $\beta_{age} = 0.0159$).

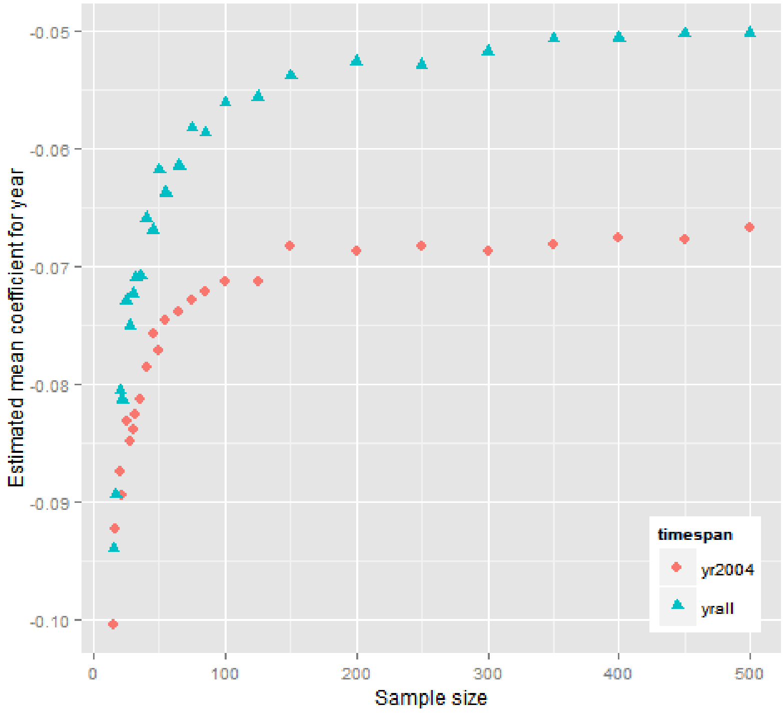

Figure 4.3: Estimated mean regression coefficient for year of diagnosis (population 1 parameter $\beta_{year} = -0.0668$ and population 2 parameter $\beta_{year} = -0.0478$).

Figures 4.4, 4.5, and 4.6 display the estimated regression coefficients and their standard errors in the manner of error bars. The error bars become shorter in an decelerated rate when sample size increases. It is also evident in these figures that the two follow-up plans have little influence in regression coefficient and standard error estimation, as the point estimator and error bars overlap. If we would like to achieve a prespecified standard error of all three coefficients below e.g. 0.01, a sample size of approximately 100 to 125 is needed. The cut off point that balances the precision and cost might be found around a sample size of 100, since the estimation perfomance improves much slower beyond size 100.

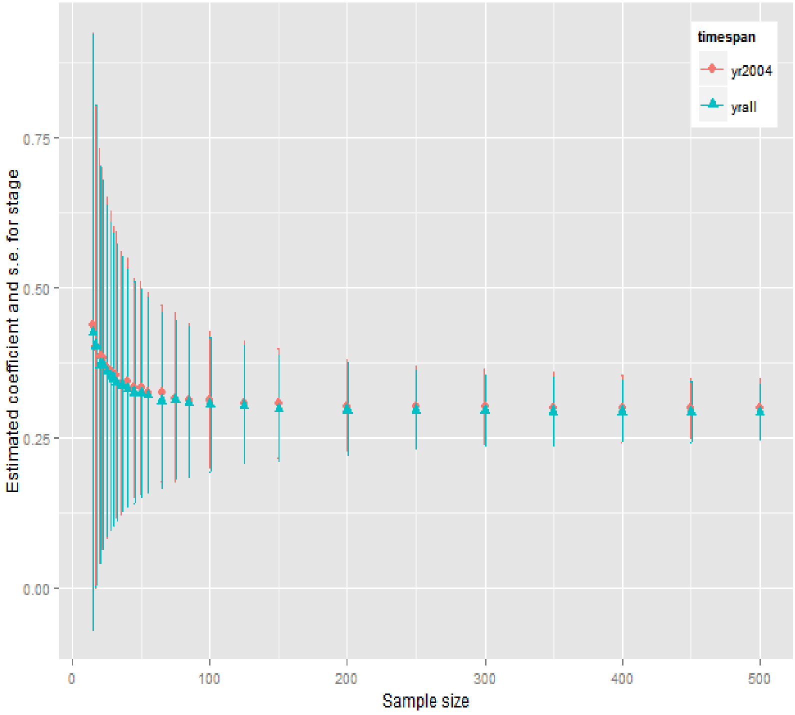

Figure 4.4: Estimated standard error of coefficient stage. The point is the mean coefficient and the error bar represents a coefficient estimate plus one standard error above the point, minus one standard error below the point. The red color represents samples from population 1 and the blue color from population 2 (population 1 $s.e._{stage} = 0.03296$ *and population 2* $s.e._{stage} = 0.0317$*).*

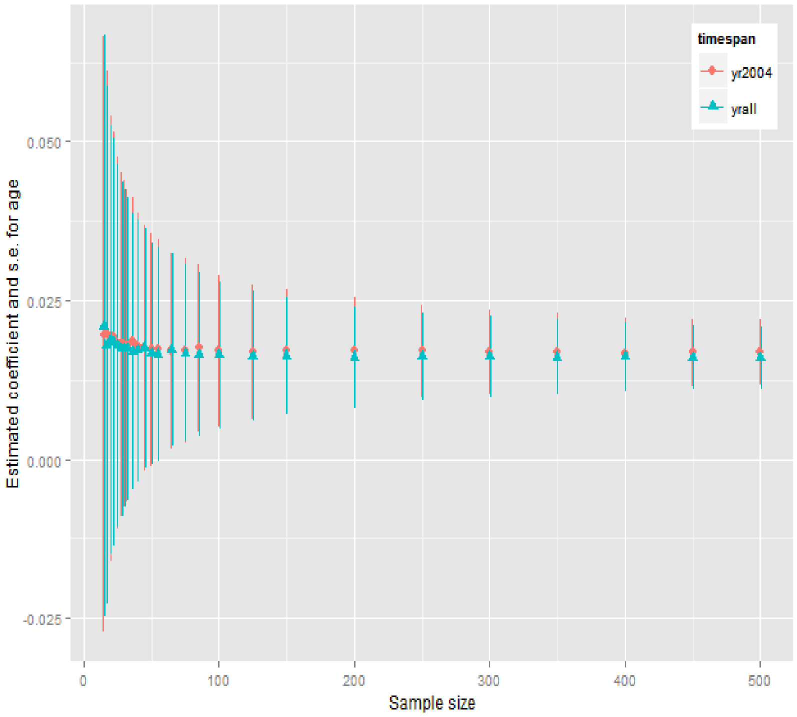

Figure 4.5: Estimated standard error of coefficient age (population 1 $s.e._{age} = 0.00331$ and population 2 $s.e._{age} = 0.0032$).

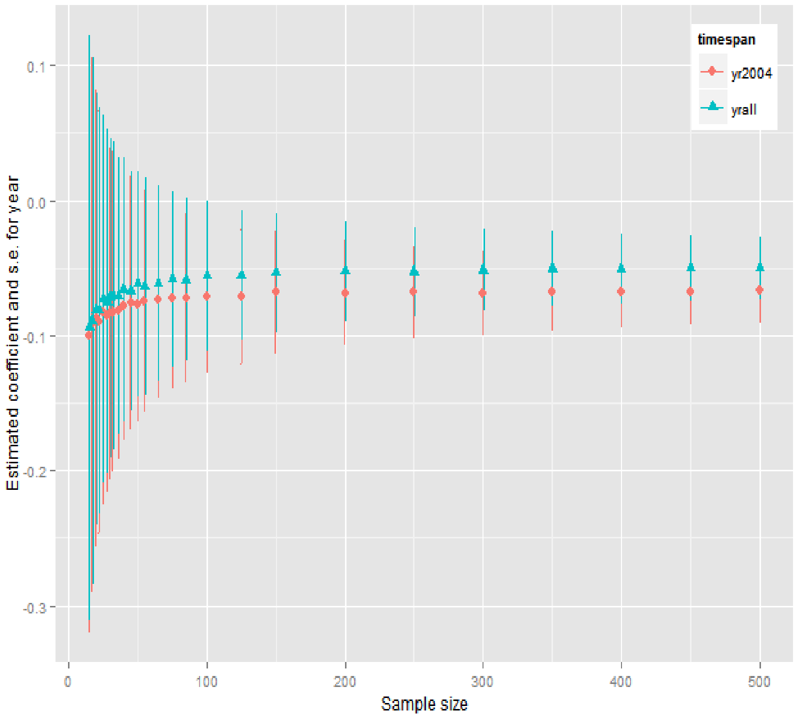

Figure 4.6: Estimated standard error of coefficient year of diagnosis (population 1 $s.e._{year} = 0.01556$ and population 2 $s.e._{year} = 0.0147$).

Chapter 5

Conclusion and Discussion

Both the survival function estimator and the Cox regression coefficient estimator are investigated. In chapter 3, we consider the inferential characteristics of the Kaplan-Meier estimator. The performance of the estimator has major improvement in respect to the MSE (mainly due to the variance) within sample sizes of up to 100 and much less change after 100. The trend that the variance gets more drastic when survival probability decreases is rather interesting, which may be accounted for by the censoring and the decreasing information with lower survival probability as discussed before. This trend may also be caused by another key factor - the simulation design. The simulation design in the thesis, is crude to some degrees, which specific needs for percentage of censoring in samples are not guaranteed. Since the simulation process is random, the degree of censoring in each sample drawn from population 1 differs, which may explain the large variance at low survival probability points where censored observations are mainly found in population 1. If every sample from population 1 has 8% censoring when simulating, we may expect the changing result, although the simulation design explain little about the trend in population 2 where there is no censoring.

In chapter 4, we consider the Cox regression coefficient estimation. The coefficient and its precision estimates show concave-down curves, which confirm the presumption of a cut-off sample size for some specified precision. The coefficient estimates of samll sample sizes, compared with those of large sample sizes, indicate overestimation of death risk when sample size is small as with an additional increase of one of the covairates level, the hazards increase more than the hazards in population. It is likely that the degree of overestimation could be a supplement measurement to decide the sample size when considering the regression models, but more thoughts are required to set up the threshold for overestimations or possible underestimations. If the standard error of the coefficient is controlled at 0.01, we may conclude that the cutoff point is around sample size of 100, which is consistent with the conclusion in the KM estimation. A very small effect of follow up time is witnessed and more meaningful results might be discovered in longer follow-up designs. Another strategy of simulation design, when simulating from population with censoring, may improve the result about the effect of follow-up.

Through demonstrating with practical examples, I show how to use simulation technique to estimate parameters, precision of the estimation, and calculate sample size. Though involving intensive computing, simulation approaches are applicable to any data-generating model, and statistical test. The flexibility of simulation enables researchers to estimate the sample size required in the complex medical study design, which might be not available in conventional sample size determination. Arnold et al. (2011) stated in their paper that it is common to examine the treatment effect in clinical trials. With simulation technique, we can determine the sample size required for detecting interaction effect at some significance level. Stahl and Landau (2013) claim that the simulation requires statisticians to clearly state the analysis procedures, which encourages investigators to be more realistic and more cautious about modelling and estimation.

In this thesis, I only consider the MSE and standard error as measurements of precision. In the Kaplan-Meier method simulation, the choice of point estimation (the quartiles survival probability associated survival time) may be improved if survival probability is compared rather than survival time, as the probability has better generality. It might be more reasonable to find the quartiles survival probability points associated survival times using the population parameter, and when we have samples, we search for the approximately corresponding survival probability using the population survival time. The sample probability will be compared to the population probability. In terms of standard error, we may use more statistical measurements such as confidence intervals to draw more reliable conclusions. And more work can be done by power analysis in determining sample size, as specifying the probability that a particular estimate will be statistically significant is also typically adopted. Due to the real data case, the results are only applicable to this cancer study.

Appendix A

Appendix

R code for simulation

```
dat=read.csv("popdata.csv",header=TRUE)
###diagnosed during 1.Jan.1990-1.Jan.1999; The end of the study 1-Jan-2004
#identify the censored observations
censor=function(data){
  n=dim(data)[1]
  status=rep(NA,n)
  ind=rep(0,n)
  ind=data$diag_yr+(data$diag_mn+data$time)%/%12
  for(i in 1:n){
    if(ind[i]>=104){
      status[i]=0
    }
    else{
      status[i]=1
    }
  }
  return(status)
}
death1=censor(data=dat)    #death indicator
cdat1=cbind(dat,death1)    #create new dataset with indicator

#modify the survival time of censored observations according to the
#study time span
surtime=function(data){
  n=dim(data)[1]
  for(i in 1:n){
    if(data$death1[i]==0){
      data$time[i]=(104-data$diag_yr[i])*12+1-data$diag_mn[i]
    }
  }
  return(data$time)
}
stime1=surtime(data=cdat1)
#my dataset to do further analysis
mydata1=data.frame(cdat1$id,cdat1$stage,cdat1$diag_yr,cdat1$age,
                   stime1,death1)
colnames(mydata1)=c("id","stage","year","age","time","death")

###detecting the longest time (in years) the patients survive
ls=function(data){
  n=dim(data)[1]
  st=rep(0,n)
  for(i in 1:n){
    st[i]=(data$diag_mn[i]+data$time[i])/12+data$diag_yr[i]-100
  }
  return(max(st))
}
ls(dat)
###The last survivor lived to June 2008

###Dataset 2
#The end of the study 1-Jan-2009; no censoring
death2=rep(1,994)
mydata2=data.frame(dat$id,dat$stage,dat$diag_yr,dat$age,dat$time,death2)
colnames(mydata2)=c("id","stage","year","age","time","death")

###
install.packages("survival")
library(survival)
###Kaplan-Meier
#population
my.surv1=survfit(Surv(time,death)~1,data=mydata1)
my.surv1
quantile(my.surv1,c(0.25,0.5,0.75))
```

```
my.surv2=survfit(Surv(time,death)~1,data=mydata2)
my.surv2
quantile(my.surv2,c(0.25,0.5,0.75))

myfit=function(data){
  f=survfit(Surv(time,death)~1,data=data)
  return(f)
}

#simulation
set.seed(77)
KM=function(s,B,data){
  n=dim(data)[1]
  m=length(s)
  index=data$id
  quan=matrix(0,nrow=B,ncol=3)
  avg=matrix(0,nrow=m,ncol=3)
  va=matrix(0,nrow=m,ncol=3)
  bias=matrix(0,nrow=m,ncol=3)
  tt=matrix(0,nrow=B,ncol=3)
  mse=matrix(0,nrow=m,ncol=3)
  for(i in 1:m){
    for(j in 1:B){
      newindex=sample(index,size=s[i],replace=TRUE)
      newdata=data[newindex,]
      quan[j,]=quantile(myfit(newdata),c(0.25,0.50,0.75))$quantile
    }
    avg[i,]=apply(quan,2,mean)
    va[i,]=apply(quan,2,var)
    bias[i,]=avg[i,]-c(4.20,8.95,51.20)
    mse[i,]=bias[i,]^2+va[i,]
  }
  res=data.frame(s,bias[,1],va[,1],mse[,1],bias[,2],va[,2],mse[,2],
                 bias[,3],va[,3],mse[,3])
  colnames(res)=c("sample.size","bias75","Var75","mse75","bias50",
                  "Var50","mse50","bias25","Var25","mse25")
  return(res)
}
```

```
s=c(30,40,50,60,75,100,125,150,200,
    250,300,350,400,450,500)
km1=KM(s,B=5000,mydata1)
km2=KM(s,B=5000,mydata2)

###Cox Regression
#population
m1=coxph(Surv(time,death)~stage+age+year,data=mydata1)
m1
summary(m1)
m2=coxph(Surv(time,death)~stage+age+year,data=mydata2)
m2
summary(m2)

#simulation
set.seed(43)
Reg=function(data){
  H0=coxph(Surv(time,death)~stage+age+year,data)
  H1=coef(summary(H0))
  betas=H1[,1]
  return(betas)
}

sim=function(s,B,data){
  n=dim(data)[1]
  m=length(s)
  nc=3
  index=data$id
  reg=matrix(0,nrow=B,ncol=nc)
  beta=matrix(0,nrow=m,ncol=nc)
  se=matrix(0,nrow=m,ncol=nc)
  for(i in 1:m){
    for(j in 1:B){
      newindex=sample(index,size=s[i],replace=TRUE)
      newdata=data[newindex,]
      reg[j,]=Reg(newdata)
    }
    beta[i,]=apply(reg,2,mean)
    se[i,]=apply(reg,2,sd)
  }
  res=data.frame(s,beta[,1],se[,1],beta[,2],se[,2],beta[,3],se[,3])
  colnames(res)=c("sample.size","stage","se.stage","age","se.age",
                  "year","se.year")
  return(res)
}
s=c(15,17,20,22,25,28,30,32,36,40,45,50,55,65,75,85,100,125,150,200,
    250,300,350,400,450,500)
est1=sim(s,B=5000,data=mydata1)
est2=sim(s,B=5000,data=mydata2)
```

Bibliography

Arnold, B. F., Hogan, D. R., Colford, Jr, J. M., and Hubbard, A. E. (2011). Simulation methods to estimate design power: an overview for applied research. *BMC medical research methodology*, 11(1):94–94.

Casella, G. and Berger, R. L. (2002). *Statistical inference*. Thomson Learning, Australia, second edition.

Cox, D. (1972). Regression models and life-tables. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES B-STATISTICAL METHODOLOGY*, 34(2):187–187.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):54–75.

Fox, J. (2002). Cox proportional-hazards regression for survival data. Available online at `http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-cox-regression.pdf`.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Kardaun, O. (1983). Statistical survival analysis of male larynx-cancer patients - a case study. *Statistica Neerlandica*, 37(3):103–125.

Klein, J. P. and Moeschberger, M. L. (1997). *Survival analysis: techniques for censored and truncated data*. Springer, New York.

Lee, E. C., Whitehead, A. L., Jacques, R. M., and Julious, S. A. (2014). The statistical interpretation of pilot trials: should significance thresholds be reconsidered? *BMC medical research methodology*, 14(1):41–41.

Stahl, D. and Landau, S. (2013). Sample size and power calculations for medical studies by simulation when closed form expressions are not available. *Statistical methods in medical research*, 22(3):324–345.

Teare, M. D., Dimairo, M., Shephard, N., Hayman, A., Whitehead, A., and Walters, S. J. (2014). Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study. *Trials*, 15(1):264–264.

Zhao, W. and Li, A. X. (2011). Estimating sample size through simulations. Available online at `http://www.pharmasug.org/proceedings/2011/SP/PharmaSUG-2011-SP08.pdf`.