CONTEXT INFORMED STATISTICS IN TWO CASES: AGE STANDARDIZATION AND RISK MINIMIZATION

Zihan Lin

Thesis submitted to the University of Ottawa

in partial fulfillment of the requirements for the

Master of Health System

Telfer School of Management

Faculty of Health System Management

University of Ottawa

Acknowledgement

Thanks for pain, and thanks for pleasure.

Thanks for my supervisors, Prof. Kevin Brand and Prof. Jonathan Li. We used to meet every once a week. This was a great time investment for them, but they all endured. Both of them carefully taught me and trained me, but in different styles. I appreciate all their efforts.

Thanks for my program director, Prof. Jonathan Patrick, who was kind and straightforward to point out my mistakes.

Thanks for Annie Baylor who patiently answered all my questions about deadlines, registration, graduation, and so on.

Thanks for Nathalie Pare who helped me with networking.

Actually, thanks for everyone in the graduate office of Telfer.

Thanks for Prof. Rim Jaber who took me as her TA and taught me what to do.

Thanks for Prof. Antoine Sauré who took me as his TA and reviewed my thesis.

Thanks for Prof. Sarah Ben Amor who was my thesis examiner.

Thanks for my girlfriend who loved me and accompanied me.

Thanks for my parents who gave me a home.

Thanks for my friends Yishen, Soumya, Acacio, ...Good luck!

Contents

I Linear Approximation of Age Standardized Mortality Rate 1

1 Introduction **2**

2 Background **4**

3 Literature Review of Mortality Indicators and Age-Adjustment Methods **8**

3.1 Indirect and Direct Age-Standardization 8

3.2 Alternative Methods of Age-Adjustment 9

3.3 Cause-Specific Mortality Rate and Potential Years of Life Lost 11

4 Linear Regression Model for Standardization Factor (SF) **12**

4.1 SF Calculation and Usage 12

4.2 Data Source and Data Preparation 13

4.3 SF Histogram 14

4.4 One-Factor Linear Predictive Model for SF 15

4.5 Candidate Predictors of SF and Model Selection 19

5 Shifting Effect of SF and Choice of Reference Population **25**

5.1 Shifting Effect of SF 25

5.2 Explanation for Shifting Effect 25

5.3 Choice of Reference Population and Impact on Centered Model . . . 28

6 Model Performance for Non-HMD (Human Mortality Database) Countries **31**

6.1 Centered SF Model for HMD and Non-HMD Countries 31

6.2 Outlier Analysis 34

7 SF for Cause-Specific ASMR and PYLL and Application of Q90 37

7.1 Cause-Specific SF Linearly Predicted by Q90 37

7.2 Predicting SF for PYLL (Potential Years of Life Loss) 38

8 Discussion 42

9 Conclusion 45

II Robust Risk Management under Covariance Uncertainty 47

10 Risk Management Based on Conditional Value-at-Risk (CVaR) 48

11 Literature Review 51

11.1 Robust CVaR Minimization Model 51

11.2 A Robust Estimator for the Covariance Matrix 52

11.3 Robust Mean-Variance Model 53

12 Robust Mean-Variance Model 56

12.1 Mean-Variance Optimization 56

12.2 Robust Mean-Variance Optimization under Covariance Uncertainty . 57

13 Robust CVaR Minimization Model 60

13.1 Robust Factor-Based CVaR Minimization under Distribution Uncertainty 60

13.2 Robust Non-Factor CVaR Minimization under Distribution Uncertainty 61

13.3 Robust Non-Factor CVaR Minimization under Distribution and Covariance Uncertainty 63

14 Numerical Experiments 66

14.1 Performance Comparisons between Robust Covariance and Sample Covariance 67

14.2 Performance Comparisons between Factor-Based and Non-Factor Model 68

14.3 Discussion 69

15 Conclusion **71**

Nomenclature

ASMR Age-Standardized Mortality Rate

CMR Crude Mortality Rate

CVaR Conditional Value at Risk

HMD Human Mortality Database

PYLL Potential Years of Life Lost

Q90 90th Quantile of Population Distribution

SF Standardization Factor

VaR Value at Risk

List of Figures

1 Population distribution 15

2 SF Histogram 16

3 Population and death distribution 17

4 Log mortality rates 18

5 SF - 1 vs ΔQ90, HMD countries 20

6 Added variable plot for Q90 and D5 23

7 Comparison of model performance with/without D5, HMD countries 23

8 Shifting effect of using different reference populations 26

9 Shifting effect of centered model 29

10 Slope vs $Q90_{ref}$ 31

11 SF - 1 vs ΔQ90 for both HMD and non-HMD countries 32

12 Comparison of model performance with/without D5, HMD + non-HMD countries 33

13 Validate Q90 = 60 as a cut-off value 35

14 Population distributions of UAE and Kuwait 36

15 SF vs Q90, all cause 38

16 SF^cvs Q90, cardiovascular diseases 39

17 SF^cvs Q90, tuberculosis 39

18 SF^cvs Q90, bladder cancer 40

19 SF^cvs Q90, HIV/AIDS 40

20 PYLL SF vs Q90, HMD and non-HMD countries 41

21 Comparison of SF histogram and PYLL SF histogram 42

22 SF *vs* Q90 for HMD and non-HMD countries, grouped by territory and income class, with WHO2000 as the reference population 45

23 Goodness-of-fit of factor model 62

24 Performance test for robust covariance approach, average of the worst 5 out-of-sample CVaRs 68

25 Performance test for non-factor model, average of the worst 5 out-of-sample CVaRs 69
26 Performance test for robust covariance approach, average of the worst 6 out-of-sample CVaRs 70

List of Tables

1 CMR calculation example 6

2 Comparison of CMR and ASMR 6

3 Age-adjustment methods 10

4 SF dispersion 15

5 Comparison of residual sum of squares across candidate predictors . . 21

6 Correlation between candidate predictors 21

7 Partial F-statistics of D5 as a second predictor 22

8 Partial F-statistics of D5 as a second predictor with reference population as a covariate 24

9 Shifting effect example 25

10 SF ranking comparison 27

11 Goodness-of-fit for slope vs $Q90_{ref}$ 30

12 SF dispersion, HMD and non-HMD countries 32

13 Partial F-statistics of D5 as a second predictor 34

14 Partial F-statistics of D5 as a second predictor with reference population as a covariate 34

15 Goodness-of-fit before and after populations whose Q90 $<$ 60 are excluded. SF vs Q90 tested with data combining all 4 reference populations and all 4 calendar years 35

16 PYLL SF dispersion, HMD and non-HMD countries 41

17 Lookup table for reference populations and their Q90 and slope (year 2000) 76

Part I

Linear Approximation of Age Standardized Mortality Rate

Abstract

When faced with death counts stratified by age, analysts often calculate a crude mortality rate (CMR) as a single summary measure. This is done by simply dividing total death counts by total population counts. However, the crude mortality rate is not appropriate for comparing different populations due to the significant impact of age on mortality and the possibility of having different age structures for different populations. While a set of age-adjustment methods seeks to collapse age-specific mortality rates into a single measure that is free from the confounding effect of age structure, we focus on one of these methods called "direct age-standardization" method which summarizes and compares age-specific mortality rates by adopting a reference population. While qualitative insights in relation to age-standardization are often discussed, we seek to approximate age-standardized mortality rate of a population based on the corresponding CMR and the 90th quantile of its population distribution. This approximation is most useful when age-specific mortality data is unavailable. In addition, we provide quantitative insights related to age-standardization. We derive our model based on mathematical insights drawn from the explication of exact calculations and validate our model by using empirical data for a large number of countries under a large number of circumstances. We also extend the application of our approximation model to other age-standardized mortality indicators such as cause-specific mortality rate and potential years of life lost.

1 Introduction

Health indicators are quantitative characteristics of a population which describe its health. Health indicators are commonly grouped into mortality indicators, morbidity indicators, nutritional indicators, etc. In this paper, we will focus on mortality indicators, though our findings may provide insights for other indicators.

The total number of health events occurring in a population (e.g., the number of deaths) is useful for determining the magnitude of a public health problem. However, the absolute number of deaths in isolation is less useful for comparisons across different populations with different sizes. Here the normal step is to divide the counts by population size, expressing them in per capita terms. A widely used example of normalization is the so-called crude mortality rate (CMR) which expresses deaths in per capita terms. This is done by simply totaling up the death counts across all age groups, then totaling up the population size across the same age groups, and then finally calculating the quotient (total death counts/total population counts). Standard caution rightfully points out that totaling counts across all age groups loses information about a key risk factor, age. For example, two populations (A and B) that are subject to an identical age-specific mortality rate schedule[1] can have different crude mortality rates exclusively as an artifact of different age structures. Therefore, comparisons of crude mortality rates between populations may be misleading if the underlying age structures differ between the populations. The comparison of age-specific mortality rates, which was first introduced by William Farr in 1841, provides a detailed yet cumbersome way of assessing mortality status for populations with different age structures (Annual Report of the Registrar General, 1841). Obviously, when many age-groups and populations are considered, this method becomes taxing. A single summary measure is therefore desired. To summarize a population mortality rate schedule in a single measure while mitigating the confounding effect of age structure, a number of age-adjustment methods have been developed to ad-

[1] In the field of health indicators, counts and related terms are typically stratified by age. For example, a population would be characterized by a sequence of mortality rates, which we denote $\{m_i\}_{i \in I}$. We refer to such a sequence as a schedule. In this case, it is mortality rate schedule.

just for age structures and facilitate comparison across populations. Among many age-adjustment methods, direct age-standardization has been widely used.

Qualitative (or ordinal) insights in relation to direct age-standardization are often discussed. For example, see these text books: Clinical Epidemiology-The Essentials, Fletcher et al; Clinical Epidemiology & Evidence-based Medicine, David L.Katz; Concepts of Epidemiology, Raj Bhopal; Cancer Epidemiology: Principles and Methods, Isabel dos Santos Silva, etc. It is generally acknowledged that if population A has a disproportionately older age structure than a reference population, population A's age-standardized mortality rate (ASMR) will be lower than its CMR, and vice versa. However, such qualitative insights cannot answer the question of how much lower? Our goal is to fill this gap. Considering that this quantitative insight has been absent for decades (noting that ASMR was introduced in 1800s, we might say centuries), our analysis would be a good supplement to existing literature.

What is more, the exact calculation of ASMR requires age-specific mortality data. When such data is unavailable, there is no way to calculate ASMR. Fortunately, our model provides an alternative way to approximate ASMR even without mortality data. We believe there may be other applications that our model may inform. These will be identified in the discussion section.

In the following sections, we present our approach to find quantitative insights with regard to direct age-standardization of crude mortality rate. Later on, we will step further and apply our quantitative insights to direct age standardization of cause-specific mortality rate and potential years of life loss. Our quantitative insights may be applied to other health indicators like morbidity indicators.

2 Background

Let P_i and D_i be the population count and death count for age group i (generally $i = 1, 2, ..., 100$). Let $m_i = D_i/P_i$ be the age-specific mortality rate for age group i. Hereafter, simply called mortality rate. Recall that $\{m_i\}$ is our mortality rate schedule. In Table 1, there are two theoretical populations (A and B), each having three columns corresponding to the schedule of death counts $\{D_i\}$, population counts $\{P_i\}$, and mortality rate $\{m_i\}$. For example, for the first age group (0-34) of population A, the death count is 20 and the population count is 1000. Therefore, the mortality rate is calculated as $20/1000 = 0.02$. Total death counts and total population counts can be found in the last row. For example, population A has a total death count of 500 $(20 + 120 + 360)$. Dividing the total death count by total population count yields the crude mortality rate (CMR $= \sum_i D_i / \sum_i P_i$). For population A, this equates to $500/10000 = 0.05$. (For population B, we get 0.03). Notice that population A has a higher crude mortality rate than population B, even though their mortality rate schedules $\{m_i\}$ are identical. A qualitative insight might note that this is because population A has an older age structure compared to population B.

While the intent of this type of comparison is assessing which population is subject to the more serious force of mortality, CMR convolutes the force of mortality with age structure, and both of the two may vary across different populations. A set of "age-standardization" methods seek to collapse age-specific mortality rates in a manner that controls for age structure. We focus on "direct age-standardization" which basically examines how many deaths would have unfolded had population X's mortality rate schedule acted upon a reference population's age structure. Mathematically, let P_i' be the reference population count for age group i, then age standardized mortality rate (ASMR) is calculated as:

$$\text{ASMR} = \frac{\sum_i \overbrace{\frac{D_i}{P_i}}^{m_i} \cdot P_i'}{\sum_i P_i'}$$

Take population A for example, the numerator of the above formula can be calculated as the inner product of population A's mortality rate schedule $\{m_i\}$ and reference population count $\{P_i'\}$: 0.02×3000+0.04×3000+0.06×4000 = 420. This numerator is actually the hypothetical total death count if population A's mortality rate schedule is to act upon the reference population's age structure. The denominator is 3000+3000+ 4000 = 10000 (i.e., the total population count). Finally we divide the numerator by denominator and get population A's ASMR 420/10000 = 0.042. We repeat the same procedure for population B and unsurprisingly get the same ASMR of 0.042, as displayed in Table 2. The age-standardization is done by using the same reference population and thus keeping the age structure the same. We can assure ourselves that age structure will have no differential impact on the result from one application to the next. This is why we expect to see the same ASMR for population A and B, since population A and B have the same mortality rate schedule $\{m_i\}$.

Notably, the ratio $\frac{\text{ASMR}}{\text{CMR}}$ (as is shown in the last row of Table 2) is less than 1 if the population has an older age structure compared to the reference population, and vice versa. For example, population A has a ratio of 0.84 because its age structure is older than the reference population. Unsurprisingly, population B has a ratio of 1.4 because its age structure is younger. When a number of populations are standardized against the same reference population, the ratio $\frac{\text{ASMR}}{\text{CMR}}$ can help to compare the age structure across populations.

Now that we have introduced CMR and ASMR, we next show their relationship in nature. Let $p_i = P_i / \sum_i P_i$ be the population proportion of age group i, and $p_i' = P_i' / \sum_i P_i'$ be the reference population proportion of age group i. Now we re-express CMR and ASMR as follow (see Chiang, 1979)

$$\text{CMR} = \frac{\sum_i D_i}{\sum_i P_i} = \frac{\sum_i \frac{D_i}{P_i} \cdot P_i}{\sum_i P_i} = \sum_i \frac{D_i}{P_i} \cdot \frac{P_i}{\sum_i P_i} = \sum_i m_i \cdot p_i$$

$$\text{ASMR} = \frac{\sum_i m_i \cdot P_i'}{\sum_i P_i'} = \sum_i m_i \cdot p_i'$$

Table 1: Age-stratified population counts and death counts for population A and B. CMR is calculated based on these information. Notice that population A and B have the same age-specific mortality rates (m_i) but different CMRs.

Age Group(i)	Population A		
	Death Count (D_i)	Population Count (P_i)	Mortality Rate (m_i)
0-34	20	1000	0.02
35-64	120	3000	0.04
>65	360	6000	0.06
Total	500	10000	0.05 (CMR)

Age Group(i)	Population B		
	Death Count (D_i)	Population Count (P_i)	Mortality Rate (m_i)
0-34	120	6000	0.02
35-64	120	3000	0.04
>65	60	1000	0.06
Total	300	10000	0.03(CMR)

Age Group(i)	Reference Population Count (P_i')
0-34	3000
35-64	3000
>65	4000
Total	10000

Table 2: While population A and B have the same age-specific mortality rates, CMR suggests that population A is exposed to higher mortality risk. In contrast, ASMR helps to make a fair comparison of the mortality risks.

Summary measures	Population A	Population B
CMR	0.05	0.03
ASMR	0.042	0.042
ASMR/CMR	0.84	1.4

It is clear that CMR and ASMR both can be expressed as a weighted average of the same mortality rate schedule $\{m_i\}$ with only one difference, namely the weights. For CMR the weights are the study population proportions $\{p_i\}$; for ASMR the weights are the reference population proportions $\{p_i'\}$. In this way, CMR can be considered as a special case of ASMR, that is, when a population itself is used as the reference.

Given that CMR is easier to calculate and that ASMR is more useful for comparison, we speculate whether ASMR can be calculated as the product of CMR and a 'standardization factor', that is, ASMR = SF $\times$ CMR. We refer to SF as the standardization factor. Its purpose as a multiplier is to convert a CMR into an ASMR. With SF so defined, our goal is to approximate SF. We will introduce the linear

approximation of SF in section 4.

3 Literature Review of Mortality Indicators and Age-Adjustment Methods

Various mortality indicators and age-adjusted methods have been proposed and reviewed by various authors (Chiang, 1979; Breslow and Day, 1980; Inskip, et al., 1983; Ahmad, et al., 2001). Table 3 contains some of the mortality indicators and age-adjustment methods found in the literature. Formulas for the standard errors associated with each method have been discussed by Chiang and Keyfitz (Chiang, 1961; Keyfitz, 1966).

3.1 Indirect and Direct Age-Standardization

In 1853, William Farr applied the age-specific death rates of the reference population to study populations and then obtained the first indirectly standardized mortality rates (Annual Report of the Registrar General for England and Wales, 1853). Indirect age-standardization and its corresponding index, the Standardized Mortality Ratio[2](SMR), were the only age-adjustment methods until the direct age-standardization and its corresponding index, the Comparative Mortality Figure (CMF)[3], were introduced in the Annual Report of the Registrar General for England and Wales (1883). For more details about SMR and CMF, readers are directed to Breslow and Day (1980).

Direct and indirect standardization are the most commonly used techniques for summarizing rates and comparing populations. Both methods have advantages and disadvantages under different situations. The direct method is advocated because it preserves consistency across the populations, that being said, if the age-specific rate of study population A is greater than the corresponding rate of study population B, then the CMF and directly-standardized rate in A will be greater than in B, irrespective

[2]SMR is the ratio of the number of observed deaths in a study population to the number that would be expected if the study population had the same age-specific mortality rates as that of the reference population.

[3]CMF is the ratio of directly age-standardized mortality rate of a study population to the crude mortality rate of the reference population.

of the standard population employed (Inskip, et al., 1983). The advantage of the indirect method is that it does not require the age-specific death counts of the study population. Therefore, the indirect method can sometimes be used when the direct method cannot. In addition, the indirect method has the advantage of a low standard error (Inskip et al., 1983).

3.2 Alternative Methods of Age-Adjustment

- Life table death rate (Brownlee, 1922). Proposed by Brownlee in 1922, the Life-Table Death Rate (LTDR) is also a weighted average of age-specific mortality rates. However, unlike CMR or ASMR, LTDR does not depend on population distribution of either the study population or reference population; instead, its weight is based on the proportion of life time spent across age groups by the life table population. calculated as the inverse of the expectation of life obtained from a life table. LTDR can also be interpreted as the inverse of life expectation at age 0 (Chiang, 1979).

- Equivalent average death rate (Yule, 1934). Equivalent Average Death Rate (EADR) is another weighted average of age-specific mortality rate, where weights are proportionate to age interval length. Notice that the last age interval is open, a proper upper limit needs to be specified such that the mortality rate in the last age interval is not overweighted (Chiang, 1979).

- Relative mortality index (Liddell, 1943). Recall the formula of CMR and replace m_i with $\frac{m_i}{m_i'}$ (ratio of study population's age-specific mortality rates to the counterpart of reference population), we obtain the Relative Mortality Index (MRI). An easy rearrangement reveals that MRI can be computed without knowledge of the study population's age-specific population counts (Chiang, 1979).

- Relative mortality index (Yerushalmy, 1951). Yerushalmy proposed this variation of MRI where the weights are lengths of age intervals.

Table 3: Age-adjustment methods

Title	Formula	Reference
Crude mortality rate	$\sum \frac{P_i}{P} m_i$	Linder, F. E. and Grove, R. D. (1943)
Direct age-standardization	$\sum \frac{P_i'}{P'} m_i$	The Registrar General's Statistical Reviews of England and Wales, 1934
Comparative mortality rate	$\sum \frac{1}{2} \left(\frac{P_i}{P} + \frac{P_i'}{P'} \right) m_i$	Statistical methods in cancer research, Breslow and Day, 1980, Chapter 3
Indirect age-standardization	$\frac{D'}{P'} \frac{D}{\sum m_i' P_i}$	The Resistrar General's Decennial Supplement, England and Wales, 1921, Part III
Life table death rate (Brownlee)	$\sum \frac{L_i}{L} m_i$	Brownlee, J., 1913; 1922
Equivalent average death rate (Yule)	$\sum \frac{N_i}{N} m_i$	Yule, G. U., 1934
Relative mortality index (Liddell)	$\sum \frac{P_i}{P} \frac{m_i}{m_i'}$	Linder, F.E. and Grove, R. D., 1943
Relative mortality index (Yerushalmy)	$\sum \frac{N_i}{N} \frac{m_i}{m_i'}$	Yerushalmy, J., 1951

Note: we use uppercase letters to denote absolute counts; lowercase letters to denote rates; prime sign to denote reference.
P_i = population count in age group i for study population; $P = \sum P_i$
P_i' = population count in age group i for reference population; $P' = \sum P_i'$
D_i = death count in age group i for study population; $D = \sum D_i$
D_i' = population count in age group i for reference population; $D' = \sum D_i'$
$m_i = \frac{D_i}{P_i}$ = mortality rate in age group i for study population
$m_i' = \frac{D_i'}{P_i'}$ = mortality rate in age group i for reference population
L_i = number of years spent in the ith age interval by a life table population, $L = \sum L_i$, and L_i/L is the proportion of life time spent in the ith age interval
N_i = length of the ith age interval, $N = \sum N_i$

3.3 Cause-Specific Mortality Rate and Potential Years of Life Lost

We will be exploring extensions of our insights for two related measures. The first one is cause-specific mortality rate, which is essentially the same with the all-cause mortality rate. For example, cause-specific mortality rate can be used to study cancer mortality (Cuzic et al., 1994; Reulen et al., 2010). In order to compare cause-specific mortality between populations, age-standardization is used (Moss et al., 1991; Adjuik et al., 2006). The second one is potential years of life lost (PYLL), which is another mortality indicator (Dempsey, 1947; Dickinson, 1948). PYLL counts deaths weighted by a penalty that measures what is lost when people die. Mathematically, PYLL $= \sum D_i w_i$, where D_i is the number of deaths[4] in the ith age group and w_i is the corresponding penalty[5]. There are analogous steps for exploring PYLL in per capita terms (i.e., a crude PYLL rate) and age-standardization. Haenszel was the first to propose a standardized PYLL, analogous to direct standardization of mortality rates (Haenszel, 1950).

In the following sections, we will first explore insights for age-standardized mortality rate and extend our insights to the age-standardization of cause-specific mortality and PYLL later on.

[4]Cause-specific deaths are often used so that PYLL helps to compare the impact of different cause of death within a population.

[5]Typically, w_i is the number of years of life lost by a person who dies in the ith age group. For example, for people who die at the age of 60, the penalty can be $w_{60} = 75 - 60 = 15$. Note that 75 is an arbitrary threshold, and if people die after this age, there will be no penalty.

4 Linear Regression Model for Standardization Factor (SF)

In this section, we will further examine the standardization factor (SF) and illustrate how SF can be used. We will propose a linear regression model for predicting SF from an easily accessible attribute called Q90.

4.1 SF Calculation and Usage

At the end of Section 2, the standardization factor (SF) was defined as $\text{SF} = \frac{\text{ASMR}}{\text{CMR}}$. The calculation of SF is straightforward, for example, in Table 2 we already have CMR and ASMR and we only need to divide ASMR by CMR to get SF. Here we explain the use of SF from various points of view.

Recall that SF can be used to compare age structures between study populations as was discussed in Section 2.

Another use of SF would be if two populations have the same mortality rate schedules, then comparing CMR is equivalent to comparing the reciprocal of SF. This is because under the assumption of identical mortality rate schedule, ASMR will be the same for the two populations. Therefore, comparing CMR is equivalent to comparing the ratio of CMR to ASMR, i.e. $\frac{1}{\text{SF}}$. For example, from Table 1 and Table 2 we know that population A and B have the same mortality rate schedules and their SF's are available. Then we have $\frac{\text{CMR}_A}{\text{CMR}_B} = \frac{\text{SF}_B}{\text{SF}_A} = \frac{1.4}{0.84} = \frac{5}{3}$.

The main use of SF allows an estimate of the ASMR to be obtained from the CMR

$$\text{ASMR} = \text{SF} \times \text{CMR}$$

For example, from Table 1 and Table 2we know population A has a CMR of 0.05 and a SF of 0.84 and we can easily derive population A's ASMR as their product: $0.05 \times 0.84 = 0.042$.

Ideally, we would have a simple linear regression model $\text{SF} = a + bx$, such that

$$\text{ASMR} = (a + bx) \times \text{CMR}$$

where a and b would be fixed parameters and x would be a predictor. There are several possible predictors. In the following subsections, we aim at finding a good x.

4.2 Data Source and Data Preparation

We base our analysis on three data sources, namely the Human Mortality Database (HMD)[6], the World Population Prospects (WPP)[7], and the Global Health Estimates (GHE)[8]. The HMD includes 39 countries, most of them are developed countries. Some data can date back to as early as 1676 (Sweden cohort data), while some data started as late as 1991 (e.g., Germany cohort data, probably because of territory change). Most recent data was from year 2015. We used data from year 2000 in order to make sure that all countries' data are available. To increase the accuracy of our analysis, we used data stratified by one-year interval (starting from age 0 up to age 110 with an increment of 1). We also collapsed the last 10 intervals, namely [100, 110). This action may only have tiny impact on our results since the population size after the age of 100 is small.

For stress-testing our model and other purposes, as we will see later, we also considered countries that are not in HMD database. We call these countries non-HMD countries. By considering these countries, we covered more than 160 countries in total. For non-HMD countries, we obtained their population data from WPP; we obtained their death data from GHE. Note that GHE only contains death data after year 2000. Unlike HMD where data is stratified by one-year age groups, these two datasets use five-year age groups. WPP uses the following age groups: 0-4, 5-9,..., 80-84, 85-89, 90-94, 95-99, and 100+, while GHE uses slightly different age groups:

[6]For more details of the HMD database, please refer to HMD website "http://www.mortality.org/" and explore their "Explanantory Notes".

[7]World Population Prospects: The 2015 Revision (United Nations, Department of Economic and Social Affairs)

[8]Global Health Estimates 2015 (World Health Organization)

0-28 days, 1-11 months, 1-4, 5-9,..., 80-84, and 85+. To align the two datasets, for WPP data, we aggregated all age groups after age 85; for GHE data, we aggregated all age groups before age 4. We ended up with the following age groups: 0-4, 5-9,..., 80-84, and 85+. Since WPP and GHE have different country lists, we used their intersection which includes more than 130 non-HMD countries.

We used spline smoothing technique to estimate certain statistics (e.g., the 90th quantile). No smoothing was used elsewhere in order to not introduce errors. We acknowledged that smoothing techniques can help us get rid of extraneous fluctuations, but this advantage can be outweighed by the unwanted errors brought in, especially when one tries to smooth the death count at early age when there are sharp inflection points.

4.3 SF Histogram

Imagine that the range of SF is from 0.99 to 1.01, for example, then there is no point to spend any effort on predicting SF. By checking the histogram of SF calculated from the HMD database[9], we assure ourselves that the dispersion of SF is reasonably significant, and age-standardization can make a difference of more than 2-fold for HMD countries ($\frac{1.063}{0.438}$, see Table 4). This suggests that predicting SF is worth the effort.

Figure 1 compares WHO2000 with the population distributions of Canada and Chile[10]. Notice that WHO2000 has a tail of moderate thickness, whereas Canada's tail is thicker and Chile's thinner. WHO2000 and Chile have much thicker left tails compared to Canada. This means that WHO2000 and Chile have younger populations than Canada. In fact, WHO2000 represents the world-wide average population distribution for the year 2000, which is generally younger than those in the HMD database. Remember than an SF that is less than 1 generally means the study popu-

[9]Note that when calculating the SF for different countries, the reference population should be the same. The issue of reference population choice will be carefully discussed in Section 5. At the moment, we use WHO2000 as the reference population (Ahmad et al., 2001).

[10]For population distributions of Canada and Chile, we used exact data without smoothing technique. WHO2000 is smooth already.

Figure 1: Population distributions for WHO2000, Canada, and Chile

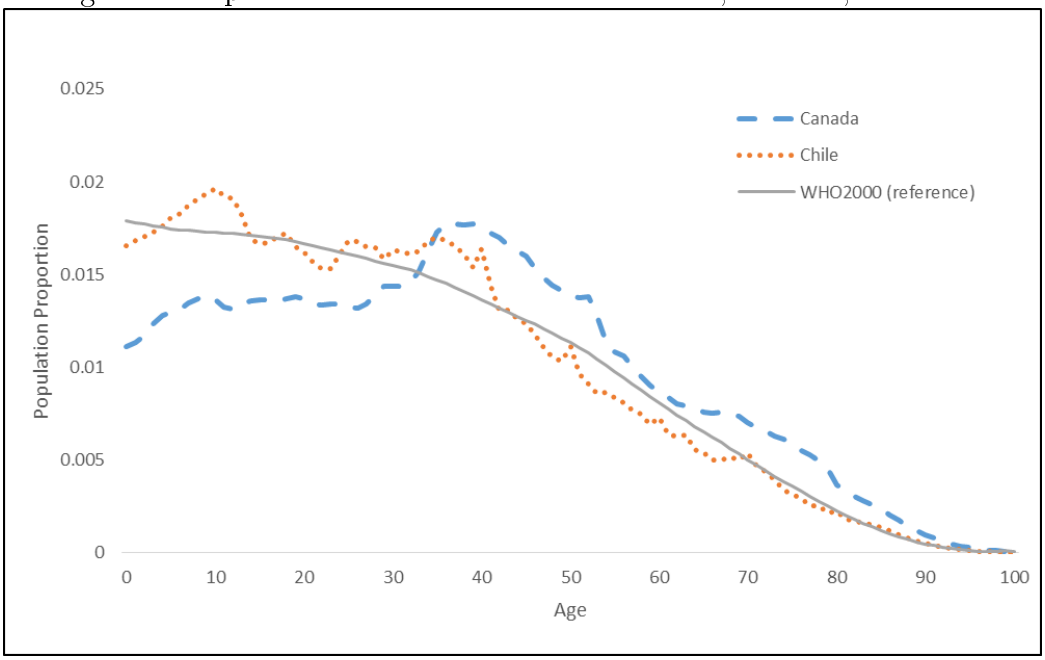


Table 4: SF dispersion

Target Populations	Minimum	Maximum	CV[11]
HMD	0.438	1.063	0.195
HMD + non-HMD	0.438	4.147	0.392
HMD + non-HMD, Q90>60	0.438	1.439	0.298

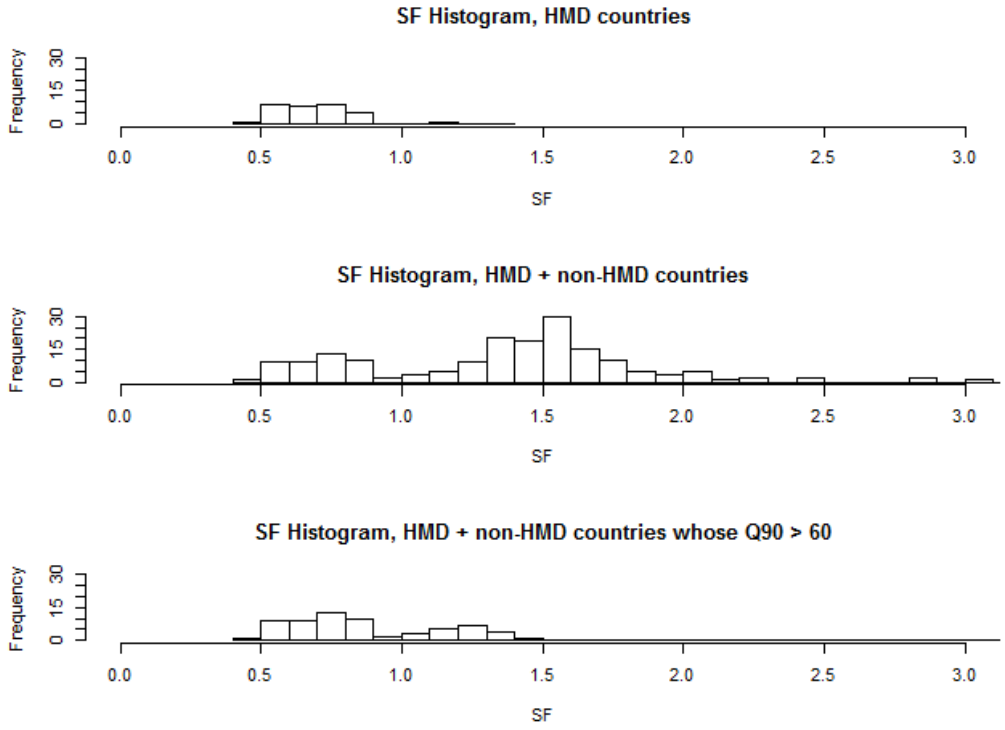
lation is older than the reference population. Given that WHO2000 is younger than most HMD countries, we expect that most HMD countries' SF would be less than 1. This expectation is supported by the histogram of SF for HMD countries (see Figure 2, and we will discuss the other two SF histograms later on.).

4.4 One-Factor Linear Predictive Model for SF

Now that the histogram demonstrates the meaning for predicting SF, we start to look for appropriate predictors. Since we considered parsimony to be a key attribute of this work, we stuck with one predictor. At the meantime, we kept an open mind to the possibility of having 2 or 3 predictors. We will discuss later on whether or not adding more predictors will improve the result.

To construct a single and powerful predictor, we look at SF in further detail.

Figure 2: SF histogram for three groups of target populations, reference population = WHO2000, year 2000



$$\text{SF} = \frac{\text{ASMR}}{\text{CMR}} = \frac{\sum_i m_i \times p_i'}{\sum_i m_i \times p_i}$$

There are three components in the definition of SF. p_i' is fixed because when we perform age-standardization, we fix a reference population. If we look at the mortality rate m_i for a number of countries, as is shown in Figure 4, we can find significant regularity (an observation that may inspire the development of mathematical models such as the Gompertz–Makeham model[12]). Compared to the plot of age-specific mortality rates, the plot of population proportions $\{p_i\}$ shows much more variation. We therefore suspect that population proportion, i.e. age structure, is responsible for the variation of SF since all other factors are almost fixed.

From figure 4, we observe that after around age 30, the mortality rate starts to grow exponentially. Given the weighted-average nature of CMR, we deduce that when the right tail of population distribution is thicker, the 'resonance' effect will

[12]See Makeham (1860) and Gompertz (1825)

Figure 3: Population and death distribution, HMD countries, year 2000

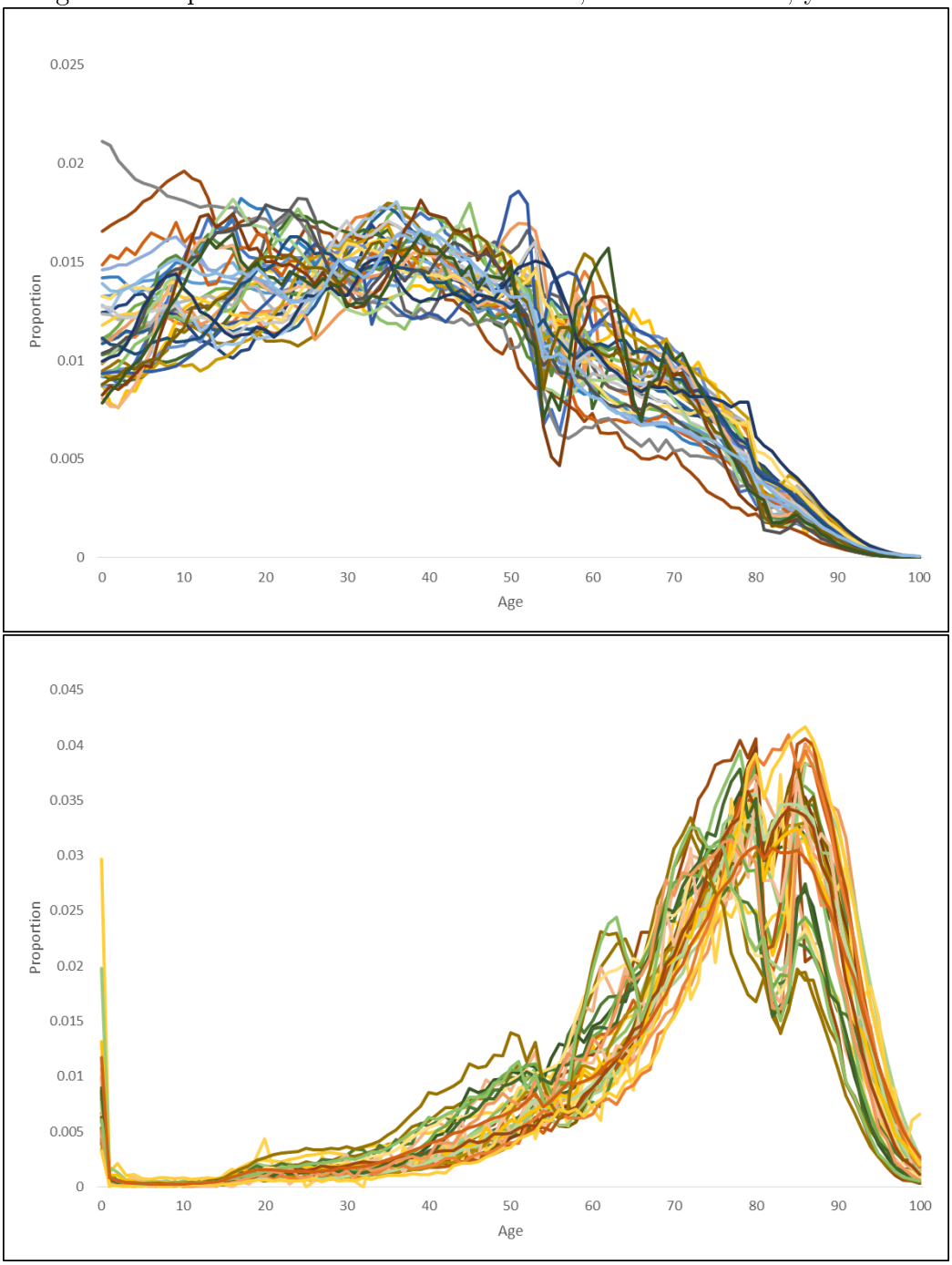
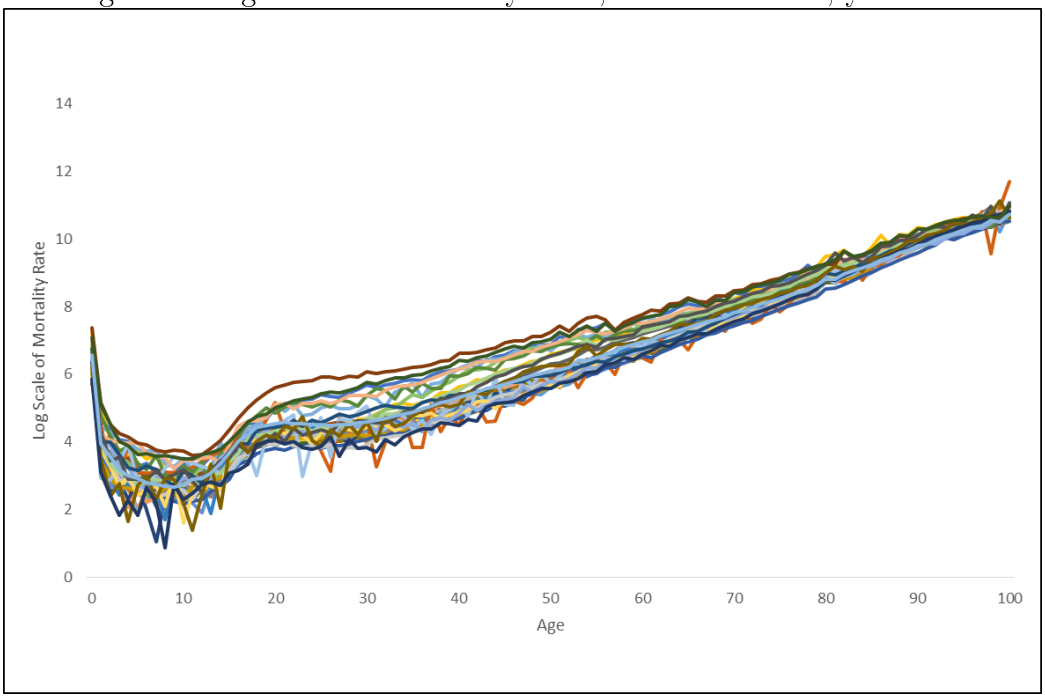


Figure 4: Logarithm of mortality rates, HMD countries, year 2000



lead to a higher CMR, and thus a lower SF. A thick right tail usually means that the population is older. Therefore, to predict SF we will look for indicators that describe the oldness of a population. One candidate indicator would be the Q90. Mathematically, given the cumulative density function of population distribution F_P, $\text{Q90} = \min\{x \mid F_P(x) \geq 0.9\} = F_P^{-1}(0.9)$. In other words, Q90 is the 90th quantile (percentile) of population distribution, representing a specific age which 90% of the population are younger than.

The linear regression model, $\text{SF} = \text{slope} \times \text{Q90} + \text{intercept}$, was tested and a strong relationship between SF and Q90 was found ($R^2 > 0.92$). However, there are two parameters in this predictive model, namely slope and intercept. For the sake of simplicity, we aim at a centered model where there is only one parameter, slope. By assuming the data point of reference, (Q90_{ref}, SF_{ref}), always lies on the regression line, we have the following centered model:

$$\text{SF} - \text{SF}_{\text{ref}} = \text{slope} \times (\text{Q90} - \text{Q90}_{\text{ref}}) = \text{slope} \times \Delta\text{Q90}$$

where $\Delta\text{Q90} = \text{Q90} - \text{Q90}_{\text{ref}}$ is the difference of Q90 between the study population and

the reference population. Notice that for the reference population ASMR = CMR, therefore $SF_{ref} = 1$. The above model becomes:

$$SF - 1 = \text{slope} \times \Delta Q90$$

In this way, it is guaranteed that the regression line always passes the origin (0,0), and the only parameter left is the slope, making the model even more parsimonious. Given the slope and the difference in Q90 between study population and reference population, we are ready to calculate SF. For example, assume that we use WHO2000 as a reference population and we know the slope is -0.05[13]. The centered model becomes

$$SF = 1 - 0.05 \times \Delta Q90$$

Suppose that population A has a Q90 that is 10 years greater than that of WHO2000, in other words $\Delta Q90 = 10$. The above equation suggests that one can estimate SF by $SF = 1 - 0.05 \times 10 = 0.5$. Remember that a smaller-than-one SF means an older population, which agrees with our assumption that population A has a higher Q90.

To see how accurate the centered model is, we test the model with HMD data. See Figure 5 for an example. We chose WHO2000 as a reference and used data of year 2000. We obtained an F-score of 446.0679, a p-value of 1.68×10^{-21}, and an R-squared of 0.9268, which suggest that the model provides a good fit[14].

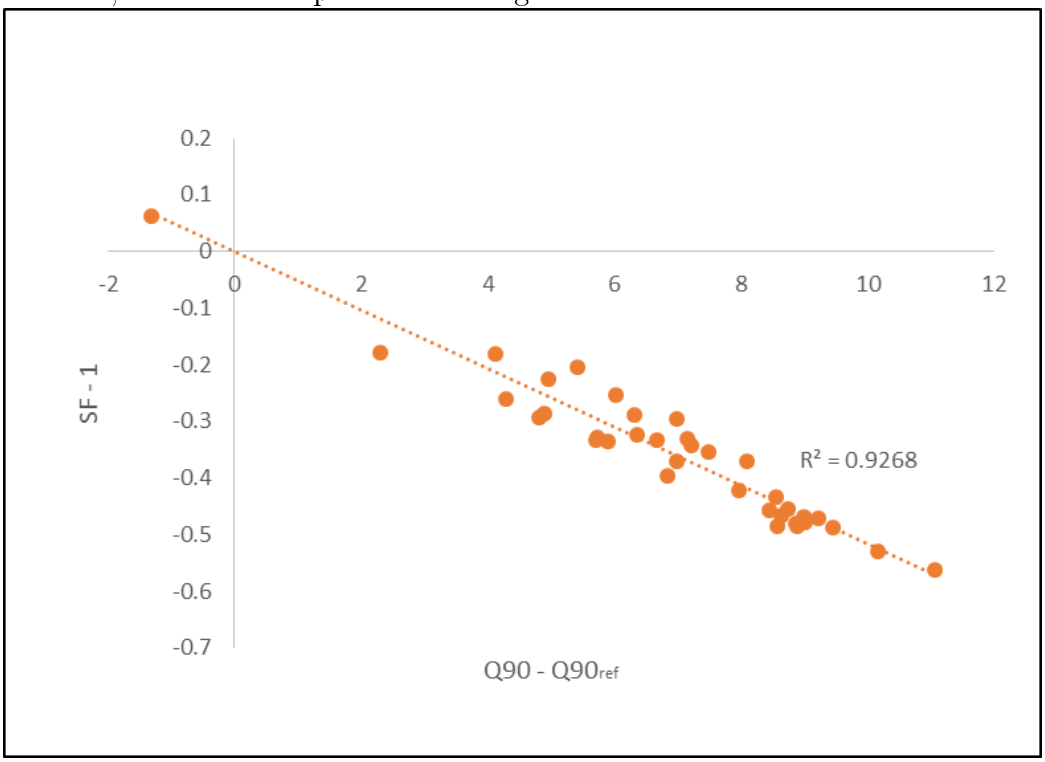
4.5 Candidate Predictors of SF and Model Selection

In this subsection, we consider a set of predictors apart from Q90. We will briefly introduce these predictors, statistically analyze the benefit of adding these predictors, rationalize our preference for one-predictor model, and suggest proper use of a one-

[13]Here we simply take the slope as granted. We will explain how the slope can be estimated in subsection 5.3.

[14]Without an intercept, R-squared may be less meaningful. Still we argue that Q90 is a good predictor of SF given the R-squared of the non-centered model is also greater than 0.92

Figure 5: SF - 1 vs ΔQ90 for HMD countries, reference population = WHO2000, year = 2000, dotted line represents the regression line



predictor model and a two-predictor model under different circumstances.

We have the following candidate predictors

- Q80 and Q95. They are variants of Q90. For example, given the cumulative density function of the population distribution F_P, Q80 $= F_P^{-1}(0.8)$, which represents the 80th quantile of the population distribution.
- P75. Given the cumulative density function of the population distribution F_P, P75 $= F_P(75)$, which represents the proportion of people who are younger than age 75.
- D5. Given the cumulative density function of the death distribution F_D, D5 $= F_D(5)$, which represents the probability that people will die before the age of 5.

We start by considering a linear regression model with only one predictor chosen from {Q80, Q90, Q95, P75, D5}, and compare their predictive power. Since another two categorical factors, year and reference population, may also have influence on SF, we used ANOVA[15] to rule out the influence of the two covariates. We then compared

[15]ANOVA assumptions were satisfied and an ANOVA was run for each candidate predictor.

Table 5: Comparison of residual sum of squares (RSS) across candidate predictors

Candidate Predictor	**RSS**
Q80	65.081
Q90	22.247
Q95	24.418
P75	40.994
D5	111.096

Table 6: Correlation between candidate predictors

	Q80	**Q90**	**Q95**	**P75**	**D5**
Q80	1	0.948	0.799	-0.777	-0.816
Q90		1	0.931	-0.905	-0.850
Q95			1	-0.984	-0.731
P75				1	0.649
D5					1

residual sum of squares (RSS) across candidate predictors and found that Q90 gave the best outcome (the smallest RSS). We also found that 'year' is an insignificant covariate (p-value = 0.229), which suggest that our model is insensitive to the change of 'year'.

When we try to improve the predictive power by bring in more predictors, the issue of multi-collinearity arises. We found that there is high correlation between these candidate predictors (see Table 6). Literature (e.g. O'brien, R. M., 2007) suggest that VIF[16] > 4 is a sign of multi-collinearity. We found that only the combination of D5 with another one predictor (two predictors in total) results in a VIF < 4. Following similar process as we did for one-predictor model, we found that the combination of Q90 and D5 results in the smallest RSS.

Now we consider a two-predictor model SF ~ Q90 + D5 and analyze whether the addition of D5 leads to improvement. We will use partial F-statistics and added variable plots and consider the following two circumstances:

- Focus on one reference population, e.g. WHO2000
- Allow multiple reference populations, i.e., treat reference population as a co-

[16]Variance Inflation Factor (VIF)

Table 7: Partial F-statistics of D5, reference population = WHO2000, year = 2000, HMD countries only

Model 1: SF ~ Q90

Model 2: SF ~ Q90 + D5

Model	**Res. Df**	**RSS**	**Df**	**Sum of Sq**	**F**	**Pr(>F)**
1	35	0.04038				
2	34	0.03839	1	0.00199	**1.762**	**0.193**

variate

As we previously discussed, 'year' is an insignificant covariate. Thus, we limited our analysis to year 2000.

When there is only one reference population (e.g. WHO2000), we found that the addition of D5 leads to insignificant model improvement of model accuracy, which is supported by partial F-statistics and p-value from Table 7. We also used added variable plots to understand the effect of D5 on SF, once the variance in SF and the variance in D5 that is explainable by Q90 has already been removed (Belsley, Kuh, and Welsch, 1980; Velleman and Welsch, 1981; Draper and Smith, 1998). For example, Figure 6 shows the value of each predictor after the other predictor is accounted for on the X-axis and the value of the SF after the other predictor is accounted for on the Y-axis, i.e., one set of residuals versus the other. By evaluating the slopes of the two charts in the grid, we can see that Q90 has much more significant influence on SF, after accounting for D5. This is because the slope of the regression line will be the coefficient of the corresponding predictor in the two-predictor model SF ~ Q90 + D5; and the less scattered the data points are, the more significant the coefficient is. In Figure 7 we plot predicted SF against true SF and compare the predicted results before (black points) and after (red points) adding D5 into the model. We conclude that there is almost no difference between the two models since black points and red points stay close to each other.

When reference population is considered as a covariate and a number of reference populations are involved, we found that the addition of D5 becomes meaningful. Table 8 shows a significant p-value for the coefficient of D5.

Figure 6: Added variable plot for Q90 and D5, reference population = WHO2000, year = 2000. We see that for the first subplot the data points fall closer to the regression line and the slope is more significant compared to the second subplot. This means that the coefficient of Q90 is more significant.

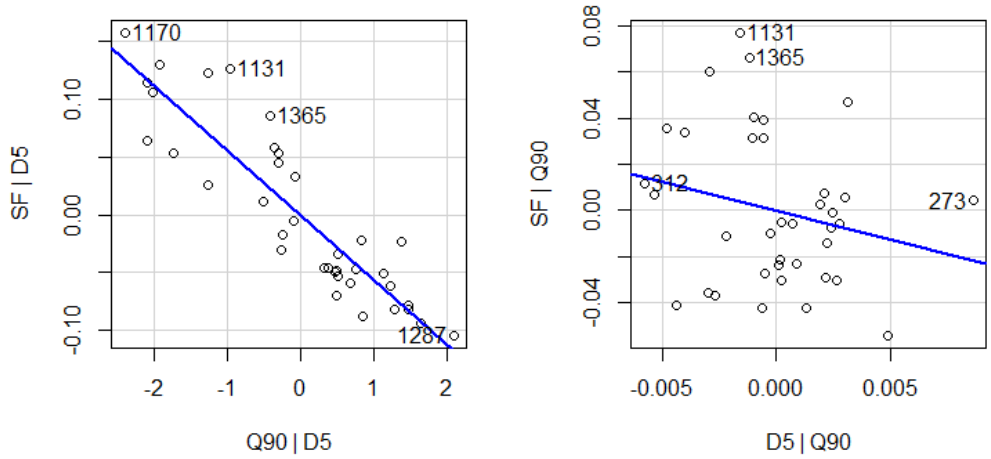


Figure 7: Comparison of model performance with/without D5, HMD countries

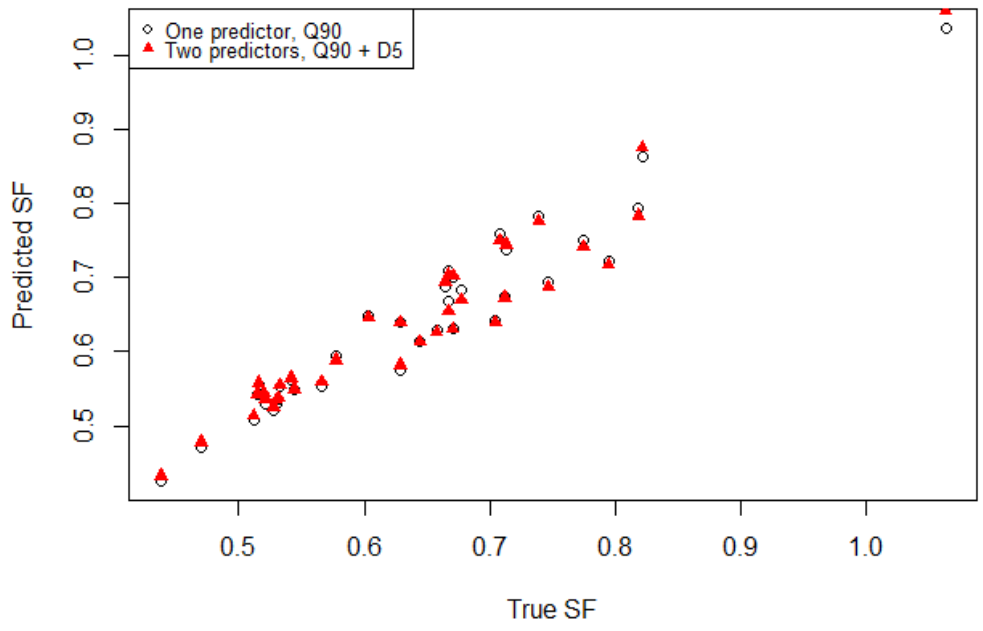


Table 8: Partial F-statistics of D5 as a second predictor with reference population as a covariate, year = 2000, HMD countries only

Model 1: SF ~ Reference + Q90

Model 2: SF ~ Reference + Q90 + D5

Model	**Res. Df**	**RSS**	**Df**	**Sum of Sq**	**F**	**Pr(>F)**
1	8618	25.092				
2	8617	24.310	1	0.782	**277.16**	**< 2.2e-16**

In conclusion, we would recommend a simple one-predictor model (SF ~ Q90) for the sake of parsimony without losing accuracy. Thus, the addition of D5 may help when multiple reference populations are involved. In fact, as we will discuss in section 6, D5 will help to some extent to improve accuracy when a larger set of countries (non-HMD countries) are considered.

Table 9: Shifting effect example

Reference	**SF (Canada, year 2000)**	**SF (Chile, year 2000)**
WHO2000	0.666	1.062
US2000	1.098	1.695

5 Shifting Effect of SF and Choice of Reference Population

It would be ideal if an ASMR is the same, regardless of the reference population used. However, it is well-recognized that the result depends upon the choice of reference population. But when it comes to how much it depends, as far as we know, has not been fully studied. For example, it is acknowledged that the choice of reference population is sometimes arbitrary, and it would be better if the reference population is not too different from study populations (Curtin, 1995; Anderson, 1998; Gillum, 2000). Here we show that different choices of reference population lead to a shifting effect. We will provide quantitative insights on this effect and its influence on SF predictive model.

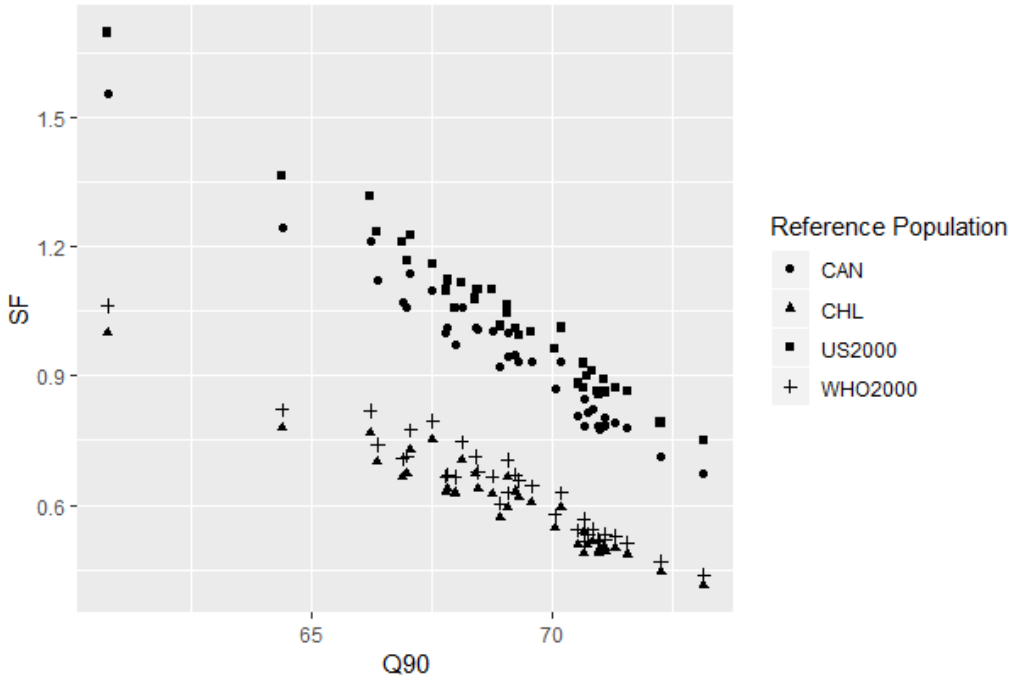
5.1 Shifting Effect of SF

Figure 8 illustrates the shifting effect of SF. For example, using WHO2000 as a reference population we have an SF of 0.666 for Canada and an SF of 1.062 for Chile. If we use US2000 as a reference population, we will have an SF of 1.098 for Canada and an SF of 1.695 for Chile. This is summarized in Table 9. Notice that the shifting effect is not "parallel". As is illustrated in Table 9, admitting US2000 brings up SF by 0.432 for Canada and 0.633 for Chile. This can also be observed after a closer look at Figure 8.

5.2 Explanation for Shifting Effect

In this subsection, we will show that the shifting effect originates from linear relationship between SF and Q90. Such linearity retains SF rankings regardless of

Figure 8: Shifting effect of using different reference populations



the reference population chosen, and when SF is plotted against Q90 the rankings will be "projected" into the shifting effect.

To see the how the shifting happens, recall that

$$\text{SF} = a + b \times \text{Q90}$$

where b is the slope and $a = 1 - \text{slope} \times \text{Q90}_{\text{ref}}$ is the intercept. For any given reference population, its corresponding regression line is determined by a and b. Data points fall close to the regression line because of the linear relationship between SF and Q90. Different choices of reference population affect a and b[17]. As a result, the regression line shifts due to changes of intercept a and slightly rotates due to changes of slope b.

Here we further show that SF ranking will not change as long as the linear relationship between SF and Q90 endures. For any two populations (A and B) and any given reference population, if $\text{Q90}_A < \text{Q90}_B$, then we have

$$\text{Q90}_A - \text{Q90}_{\text{ref}} < \text{Q90}_B - \text{Q90}_{\text{ref}}$$

[17] Clearly, different reference population results in different Q90_{ref}. In the next subsection, we will discuss how reference population affects slope.

Table 10: SF ranking slightly changes when reference population changes.

Country	**SF Ranking**	
	Reference = US2000	**Reference = WHO2000**
Sweden	1	1
Italy	2	2
Germany	3	4
France	4	9
United Kingdom	5	7
Norway	6	3
Japan	7	6
Switzerland	8	5
Belgium	9	8
Denmark	10	12
Spain	11	10
Greece	12	11
Austria	13	13
Portugal	14	14
Finland	15	15
Estonia	16	20
Hungary	17	19
Latvia	18	24
Bulgaria	19	18
Netherlands	20	16

Since 'slope' is always negative, we have

$$1 + \text{slope} \times (\text{Q90}_A - \text{Q90}_{\text{ref}}) > 1 + \text{slope} \times (\text{Q90}_B - \text{Q90}_{\text{ref}})$$

$$\text{SF}_A > \text{SF}_B$$

In other words, as long as our linear model works well, $\text{Q90}_A < \text{Q90}_B$ leads to $\text{SF}_A > \text{SF}_B$ no matter which reference population is used. As a result, SF rankings remain unchanged. Table 10 compares SF rankings when using two different reference populations. There are minor changes in SF ranking, because the linearity assumption

is not well-satisfied in real world, leading to small changes in rankings.

5.3 Choice of Reference Population and Impact on Centered Model

In the previous subsection, we mentioned that difference choices of reference population affect SF predictive model. Now that centered model is our major interest and there is only one parameter, namely slope, we will quantitatively study the influence of reference population on slope in this subsection.

When SF predictive model is centered and only the slope parameter is left, the aforementioned shifting effect becomes a rotating effect, which is shown in Figure 9. The first subplot contains four panels ordered by slope and reference Q90. For example, the first panel (second quadrant) uses US2000 as a reference population. Notice that US2000 reference has the steepest slope (-0.077) and the greatest Q90 (68.4) among all four panels. The last panel, where Chile (year 2000) is used as a reference, has the gentlest slope (-0.047) and the smallest Q90 (60.8). The difference in slope (-0.077 and -0.047) matters. Imagine ΔQ90 = 10, such a difference in slope can results in 0.3 difference in SF. We forced all four regression lines to pass through the origin (0, 0) for the sake of parsimony, but we do lose some accuracy that is tolerable (e.g., see the regression line in the first panel). If we merge the four panels, we get the second subplot where the regression lines rotate around the origin.

Our goal is to quantify the relationship between reference population and the slope. Through numerical experiment, we found a linear relationship between the slope and $Q90_{ref}$. Summary statistics (see Table 11) suggest that it is a good fit. Figure 10 visualizes such linearity, where each data point represents a reference population. For example, if we use Chile (CHL) or WHO2000 as the reference population, the slope of the centered regression model will be about -0.05. If we use Sweden (SWE) as a reference population, the slope will be around -0.09. We found that if a reference population has a higher Q90 (older age structure), then the slope of its corresponding model will be steeper.

Figure 9: Shifting effect of centered model. Regression line is centered at the origin for the sake of parsimony. Notice that shifting effect results in the changes of regression slopes.

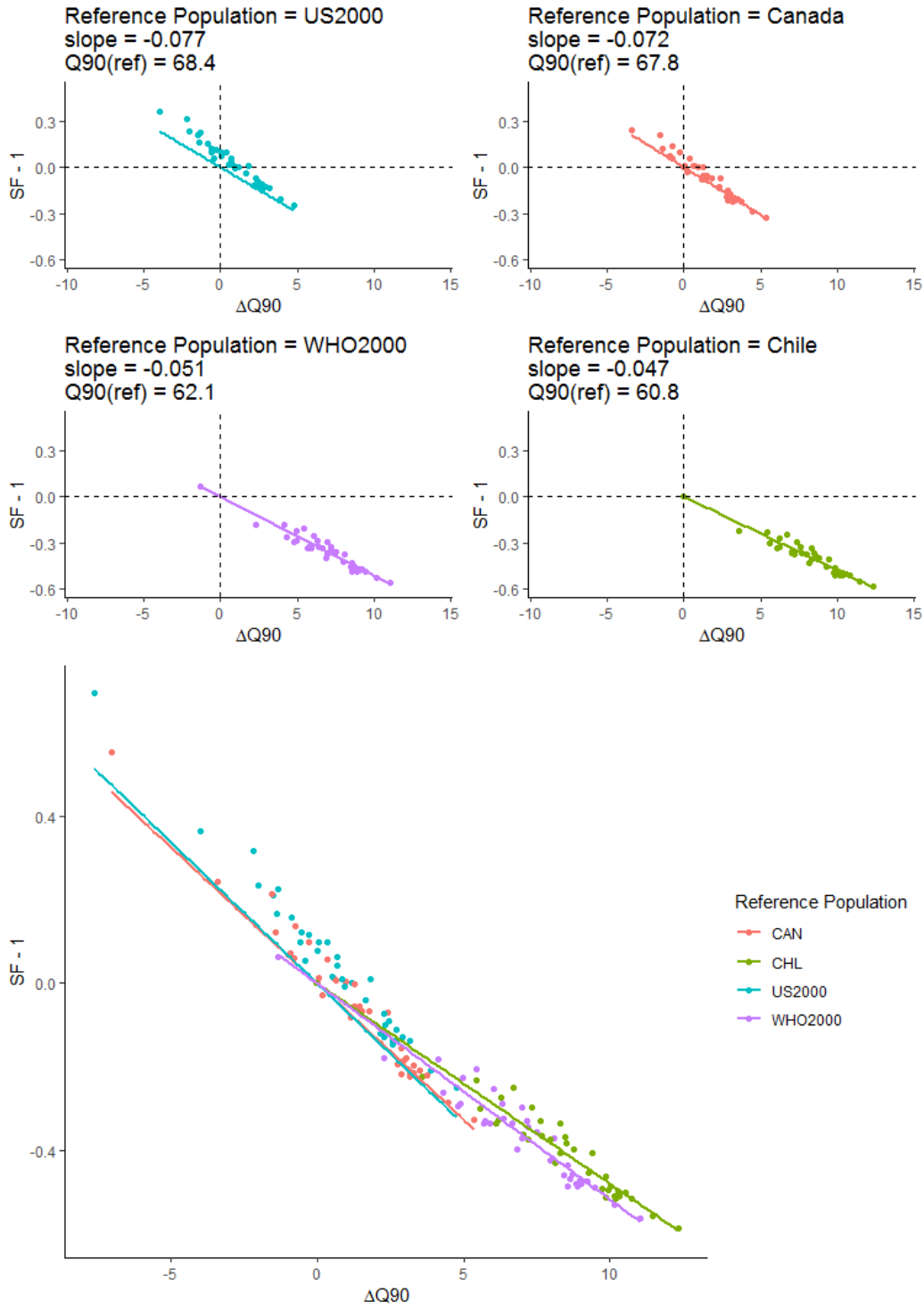


Table 11: Goodness-of-fit for slope vs $Q90_{ref}$

Coefficients:	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	0.21	0.026	7.92	1.54e-11
Q90 of reference population	-0.0041	0.00038	-10.719	<2e-16

The purpose of introducing this linear relationship between reference Q90 and slope is to facilitate the understanding of the shifting effect. Users should be careful when using this model, because it can lower the accuracy. Even though the goodness of fit for this model is satisfactory, when the estimated slope is multiplied by ΔQ90, the estimation error will be enlarged. For example, assume the study population is Canada (Q90 = 67.8) and the reference population is WHO2000 (Q90 = 62.1), then according to Table 11 we have

$$\text{SF} = 1 + \text{slope} \times \Delta\text{Q90}$$

$$= 1 + (0.21 - 0.41 \times \frac{\text{Q90}_{\text{ref}}}{100}) \times \Delta\text{Q90}$$

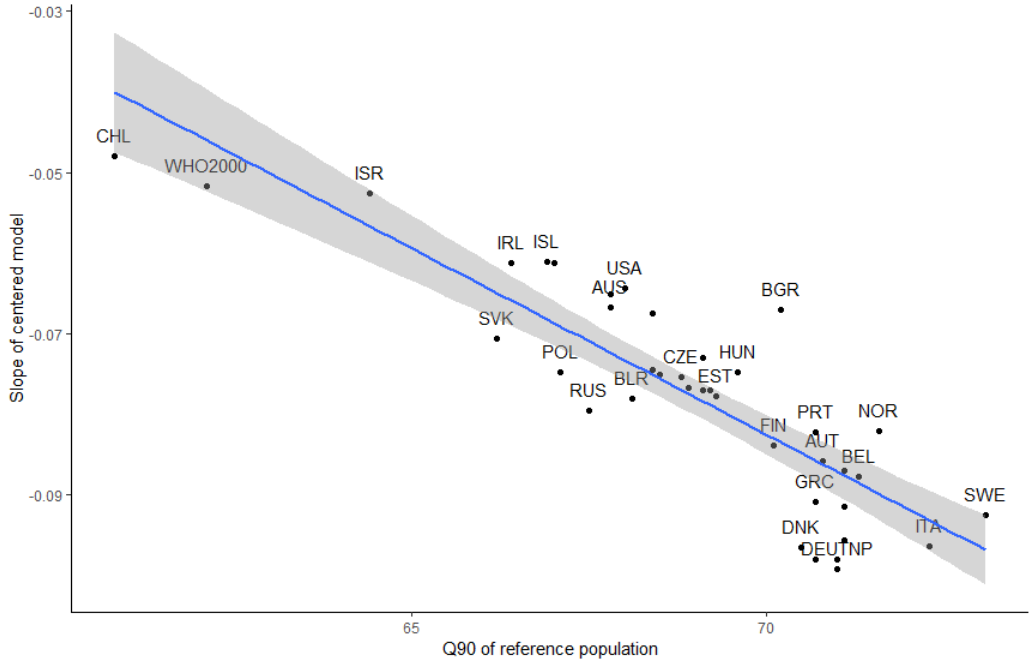
$$= 1 + (0.21 - 0.41 \times \frac{62.1}{100}) \times (67.8 - 62.1)$$

$$= 0.746$$

The exact SF should be 0.666, which means 12% discrepancy between the exact SF and the estimated SF.

In Table 17 (see Appendix), we provide a table containing a wide range of reference populations and their Q90 and slope, so that one can use our SF model without knowing reference population distribution.

Figure 10: Slope vs $Q90_{ref}$, HMD countries, year 2000



6 Model Performance for Non-HMD (Human Mortality Database) Countries

In previous sections, we discussed the linear predictive model for SF, using Q90 as a predictor, where we limited our discussion within HMD countries. As the majority of HMD countries are developed countries, such as Canada and UK, we want to stress-test our model with a wider set of countries. We will see that the linear relationship between SF and Q90 still holds for a number of countries, but there are exceptions, for which we will provide preliminary explanations. We will also suggest a simple criteria for users to do self-assessment in terms of whether or not they should use this model.

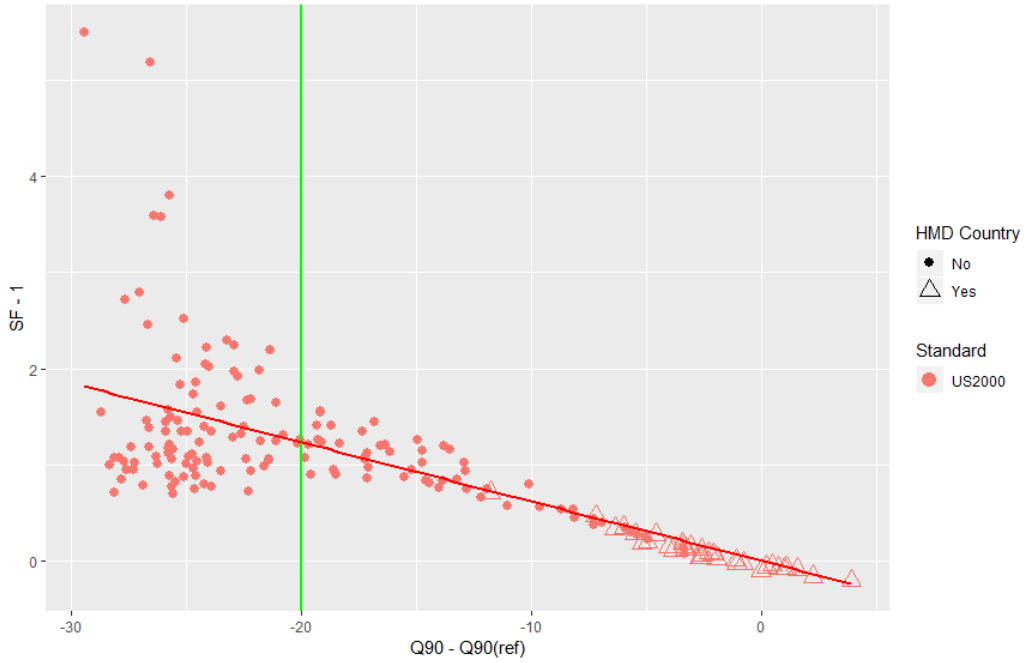
6.1 Centered SF Model for HMD and Non-HMD Countries

Since we take into account a wider set of countries, the range of SF becomes wider. As is shown in Figure 2 and Table 4, the coefficient of variation (CV) becomes higher and age standardization can make a difference of more than 8-fold ($\frac{4.147}{0.438}$). Since

Table 12: SF dispersion, HMD and non-HMD countries

Minimum SF	Maximum SF	Coefficient of Variation
0.438	4.147	0.392

Figure 11: SF - 1 ~ ΔQ90 for both HMD and non-HMD countries. Apparent outliers can be found on the left side. Notice that HMD countries all gather at the right side.

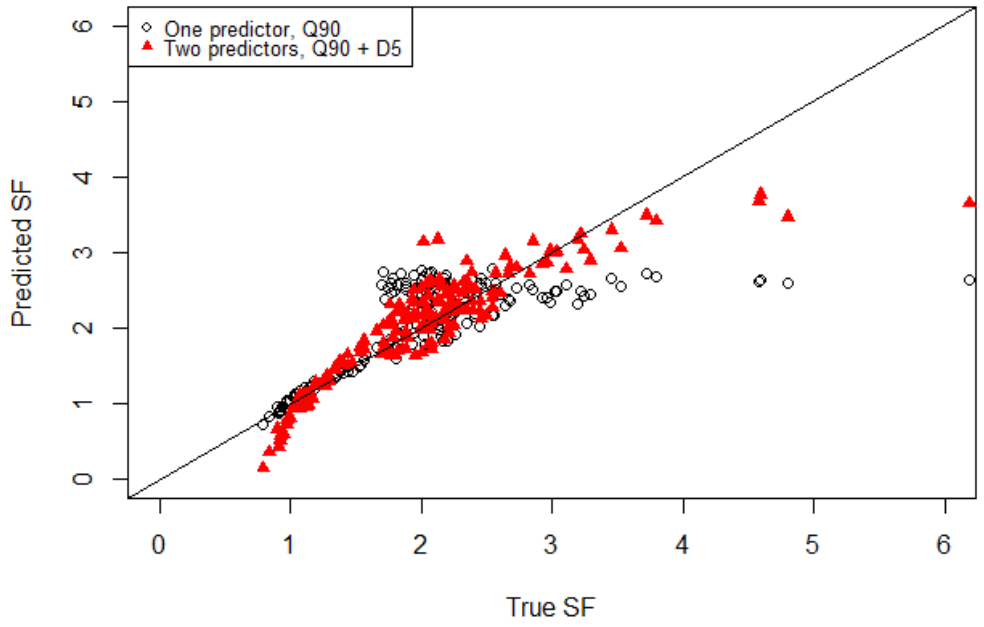


this is a reasonable amount of contrast, we believe analyzing SF for both HMD and non-HMD countries is meaningful.

In Figure 11, we used US2000 as our reference population and plotted centered SF against ΔQ90 for more than 160 countries. On the right side, the data points fall close to the regression line (the red line), but there is a 'cloud' on the left side. This means that older study populations fit well into our model, but if the study population is relatively young, whether it would fit into our model is unclear at this stage. The adjusted R-squared is 0.473, but the p-value for the slope is still significant ($<$2e-16). The green vertical line represents ΔQ90 $= -20$. Note that -20 is just a value for illustration purpose. We performed another linear regression for data points to the right of the vertical green line and we got an adjusted R-squared of 0.928. In the following subsection, we will present an intuitive and conservative criteria that gives guidance about when our approach can be sensibly used.

Recall that D5 improves predictive power when multiple reference populations are

Figure 12: Comparison of model performance with/without D5, HMD + non-HMD countries. Black line is a one-to-one line.



used (Subsection 4.5). Here we will test whether D5 will help to improve the accuracy of our model.

Firstly we fix a reference population and assess the benefit of adding D5 into the model. To visualize the influence of the addition of D5, we first use Q90 alone as a predictor and plot true SF against predicted SF. Then, we include D5 into the model and super-impose the new set of predicted SFs in the same plot. For example, in Figure 12 where reference population is US2000 and year is 2000, red points represent predicted results with D5 and black points represent predicted results without D5. The blue line is a one-to-one line on which data points would ideally lie. Clearly, red points lie closer to the one-to-one line, meaning that including D5 results in more accurate predictions. This is also supported by partial F-statistic of D5 (p-value $<$ 2.2e-16), as is summarized in Table 13.

Secondly, we allow multiple reference populations (namely WHO2000, US2000, Canada, and Chile) and find partial F-statistics of D5 still significant. This is summarized in Table 14.

Table 13: Partial F-statistics of D5 as a second predictor, reference population = US2000, year = 2000, HMD and non-HMD countries

Model 1: SF ~ Q90

Model 2: SF ~ Q90 + D5

Model	**Res. Df**	**RSS**	**Df**	**Sum of Sq**	**F**	**Pr(>F)**
1	354	129.740				
2	353	77.272	1	52.468	**239.96**	**<2.2e-16**

Table 14: Partial F-statistics of D5 as a second predictor with reference population as a covariate, year = 2000, HMD and non-HMD countries

Model 1: SF ~ Reference + Q90

Model 2: SF ~ Reference + Q90 + D5

Model	**Res. Df**	**RSS**	**Df**	**Sum of Sq**	**F**	**Pr(>F)**
1	1419	315.56				
2	1418	199.03	1	116.52	**830.15**	**<2.2e-16**

In conclusion, when multiple reference populations and non-HMD countries are taken into account, D5 will help to improve model performance at a cost of model complexity. However, since the key of this work is parsimony and people will be dealing with one reference population for most of the time, we suggest that Q90 alone is sufficient.

6.2 Outlier Analysis

In order to use the SF regression model safely, we need to know in the first place what characterize an outlier. Here we suggest that our model be reliable for countries whose Q90 $>$ 60, no matter if it is an HMD country or not. We tested this criteria for a number of circumstances. Graphical results are shown in Figure 13, where green vertical lines represent Q90 = 60. Statistical evidence in Table 15 shows that our model works well for countries whose Q90 $>$ 60. Although we tried to establish other criteria for distinguishing outliers, Q90 = 60 is the most user-friendly and most reliable one. This is because we assume users of our model have no other information except for study population proportion. Some of our additional efforts can be found in the discussion section.

Here we use an example to briefly illustrate why the model does not work well for

Figure 13: SF vs Q90, comparison across the combination of 4 reference population and 4 years. This helps to validate that Q90 = 60 as a cut-off value works for a wide range of circumstances.

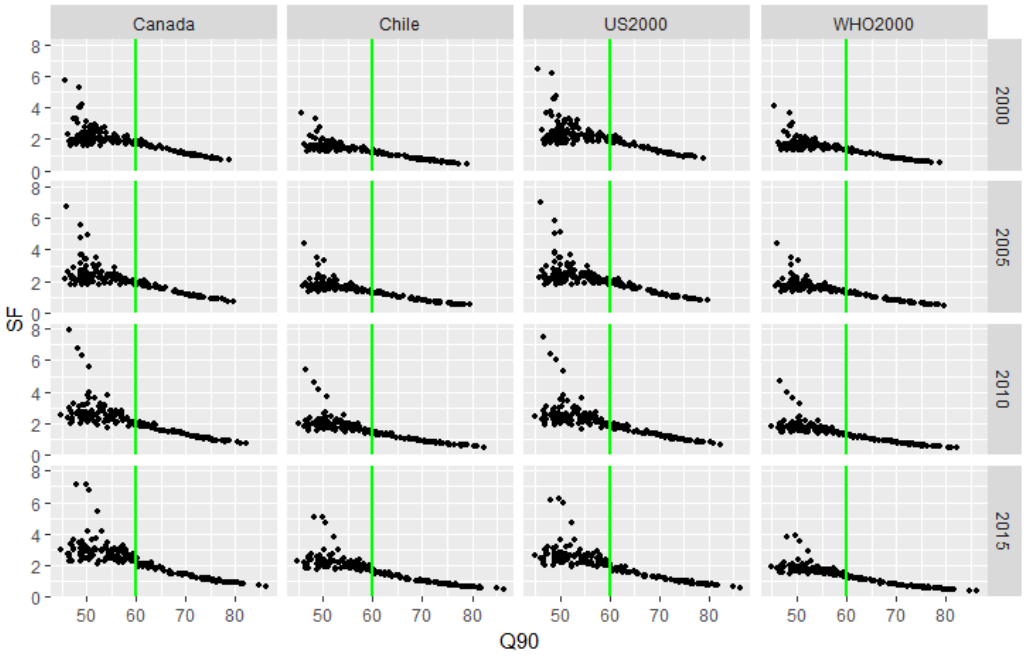
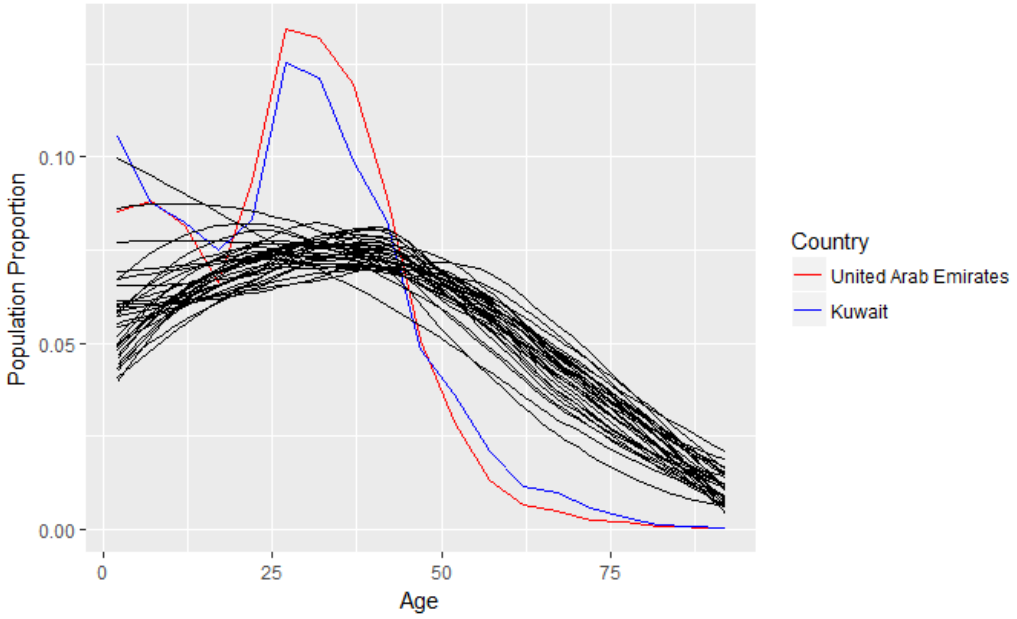


Table 15: Goodness-of-fit before and after populations whose Q90 < 60 are excluded. SF vs Q90 tested with data combining all 4 reference populations and all 4 calendar years

Cohort	Adjusted R-squared
Full cohort	0.639
Trimmed cohort (Q90>60 only)	0.926

Figure 14: United Arab Emirates and Kuwait as examples of young population distributions. Black lines represent population distributions of HMD countries.



Q90 < 60. Consider the population distributions of United Arab Emirates (UAE) and Kuwait and compare them with those of HMD countries (see Figure14). Clearly, UAE and Kuwait have very young populations (Q90 < 45) and they have disproportionally small CMRs (0.0019 for UAE and 0.0026 for Kuwait). In comparison, Canada has a CMR of 0.0071 (year 2000). ASMRs of UAE and Kuwait are thus drastically scaled up by such low CMRs, making the two countries the most apparent outliers in terms of SF (the two data points on the top of Figure 11).

7 SF for Cause-Specific ASMR and PYLL and Application of Q90

In this section, we extend the application of Q90 to cause-specific ASMR and age-standardized PYLL. We found that Q90 works well for some causes of death that share a commonality. Given cause-specific CMR and Q90 of the study population, its corresponding cause-specific ASMR is readily available as the product of cause-specific CMR and cause-specific SF, which is exactly what we do for all-cause mortality rate. As for PYLL, we argue that age-standardization of PYLL makes no big difference among study populations and therefore SF for PYLL will be unnecessary.

7.1 Cause-Specific SF Linearly Predicted by Q90

Cause-specific mortality rate is the mortality rate from a specific disease, e.g. heart disease. Cause-specific counter-part of mortality rate (m_i), CMR, and ASMR can be calculated by substituting total deaths (all-cause deaths) with deaths caused by a certain disease (cause-specific deaths). For example, cause-specific mortality rate for age group i, denoted as m_i^c, can be calculated as $m_i^c = \frac{D_i^c}{P_i}$ where D_i^c is the cause-specific death count for age group i. Similarly, we can calculate cause-specific CMR (CMR^c), cause-specific ASMR (ASMR^c), and cause-specific SF (SF^c):

$$\text{SF}^c = \frac{\text{ASMR}^c}{\text{CMR}^c} = \frac{\sum_i m_i^c \times p_i^{'}}{\sum_i m_i^c \times p_i}$$

Here we show a few examples (Figure 16 to Figure 18) where Q90 works well. We found that in these examples, cause-specific mortality rates are exponential-like, which is similar to all-cause mortality rate (Figure 15). Another example where Q90 does not work well is shown in Figure 19 where the cause-specific mortality rates are significantly different from the all-cause one.

In the cases where our model works, we can see that Q90 = 60 is still useful, though conservative, for separating outliers. However, further work needs to be done

Figure 15: SF vs Q90, all cause, year 2000, reference population = US2000

to characterize certain causes for which our model does not work well.

7.2 Predicting SF for PYLL (Potential Years of Life Loss)

Recall that

$$\text{PYLL} = \sum_i D_i \times w_i$$

where w_i is the number of years lost at the time of death in age group i. Typically, $w_i = 0$ for any age group beyond age 75. Normalize PYLL by population size and we get crude PYLL rate

$$\text{CPYLL} = \frac{\text{PYLL}}{P} = \sum_i p_i \times m_i \times w_i$$

Similarly, we can replace p_i with $p_i^{'}$ to obtain age-standardized PYLL

$$\text{ASPYLL} = \sum_i p_i^{'} \times m_i \times w_i$$

SF for PYLL will be the ratio of ASPYLL to CPYLL

Figure 16: SF^cvs Q90, cardiovascular diseases, year 2000, reference population = US2000

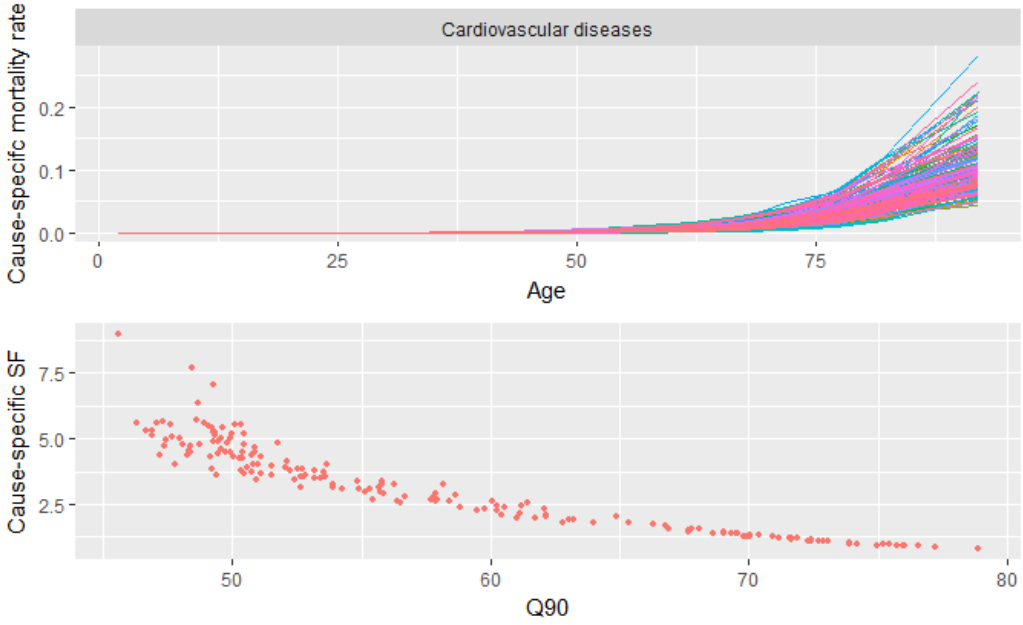


Figure 17: SF^cvs Q90, tuberculosis, year 2000, reference population = US2000

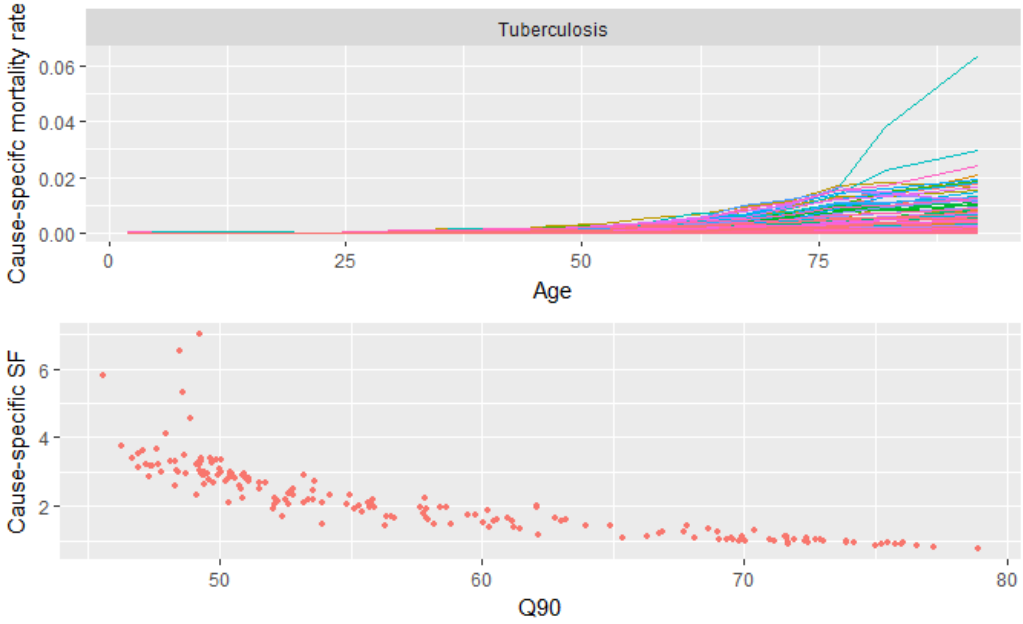


Figure 18: SF^{c}vs Q90, bladder cancer, year 2000, reference population = US2000

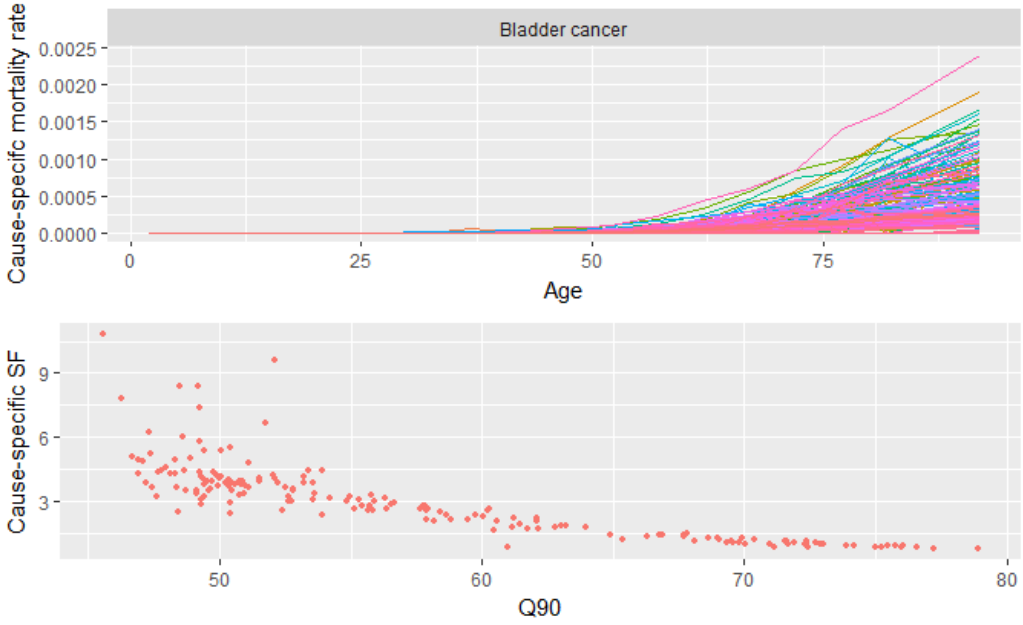


Figure 19: SF^{c}vs Q90, HIV/AIDS, year 2000, reference population = US2000

Figure 20: PYLL SF vs Q90, HMD and non-HMD countries

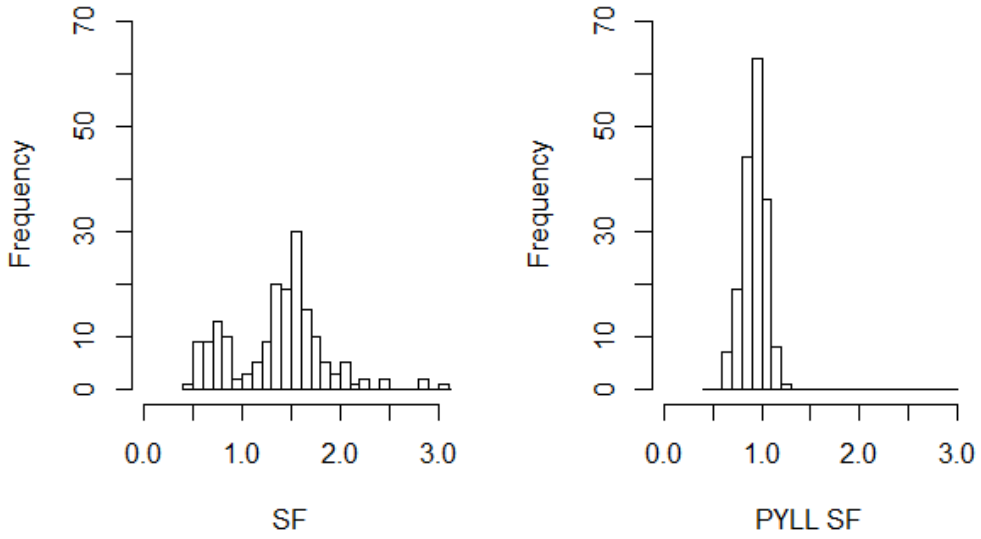
Table 16: PYLL SF dispersion, HMD and non-HMD countries

Minimum PYLL SF	**Maximum PYLL SF**	**Coefficient of Variation**
0.641	1.300	0.125

$$\text{PYLL SF} = \frac{\text{ASPYLL}}{\text{CPYLL}} = \frac{\sum_i p_i^{'} \times m_i \times w_i}{\sum_i p_i \times m_i \times w_i}$$

In Figure 20, we found the relationship between PYLL SF and Q90 is non-linear. Combining Table 16 with Figure 21, we found that the range of PYLL SF is much narrower than that of SF. This is to say, the standardization of crude PYLL rate (CPYLL) is less meaningful compared to the standardization of CMR. Given these considerations, we terminate the application of Q90 to CPYLL until more interests arise.

Figure 21: Comparison of SF histogram and PYLL SF histogram for HMD and non-HMD countries, year 2000



8 Discussion

Our model works well for HMD countries, but when it comes to non-HMD countries especially those whose Q90s are below 60, we noticed that they do not fit into the regression line very well. Therefore, we suggest users of our model look at their Q90s first. If they are below 60, users should be cautious when applying our model. In fact, we did explore factors that might distinguish those 'outlier' populations. We looked at D5, CMR, and another quantitative indicator defined as $\log(m_{80}/m_{30})$, which basically describes how steep is the slope of $\log(m)$. We also looked at territory and economic status[18]. For example, in Figure 22 we compare the relationship between SF and Q90 across different groups of countries and we observed two interesting things. Firstly, most of the country's populations were aging, making their Q90s higher in year 2015 than in year 1995. To see this, notice that light blue points, representing most recent years, usually gather at the bottom right part of each panel, while dark blue points gather at the top left part of each panel. Secondly, we noticed that when

[18] We grouped countries into four income class, according to Updated country income classifications for the World Bank's 2018 fiscal year

Q90 is above 60, the relationship between SF and Q90 is linear, while when Q90 is below 60, the data points are more scattered. In particular, high income Asian countries (middle-east countries generally belong to this group) and African countries except for South Africa have relatively low Q90 (often < 60) and have the most scattered patterns. So far, we found no criteria that is as robust and straightforward as 'if Q90<60 then caution'. By adopting this criteria, we acknowledge that we may be a little be conservative, as we can see in Figure 22, Asian countries that are in the upper middle income class, lower middle income class, and low income class do show strong linearity, despite the fact that their Q90s are below our threshold. Recall that one key feature of our work is parsimony and we assume that mortality data is unavailable. It would be counterproductive to find a criteria that is complicated and/or requires mortality data as input. Therefore, for pragmatic purpose, we stick to Q90 = 60 as a cut-off value.

Future work could be done to reveal better ways to inform people when or when not to use our model. One may calculate a Total Absolute Curvature, which is a number defined by integrating the absolute value of the curvature around a curve (J. W. Rutter, 2000), for the probability density function of the population distribution. Judging from Figure 14, we speculate that outlier countries would have more curvature. Other possibilities might include Skewness and Kurtosis of the population distribution.

In order to estimate Q90 for non-HMD countries, we had to use 5-year interval data and smoothing functions. It is possible that the estimated Q90 suffers from errors introduced by smoothing. Meanwhile, we used 5-year interval data to calculate SF for non-HMD countries, so the resulting SF is coarse. We suspect that imprecise Q90 and SF for the non-HMD countries contribute more or less to the presence of outliers. We would like to see how data of better quality (e.g., 1-year interval data) may help improve the performance of our model. Some outliers may fit well into our model given finer data.

There are two applications of our analysis that are worth discussing. One is more

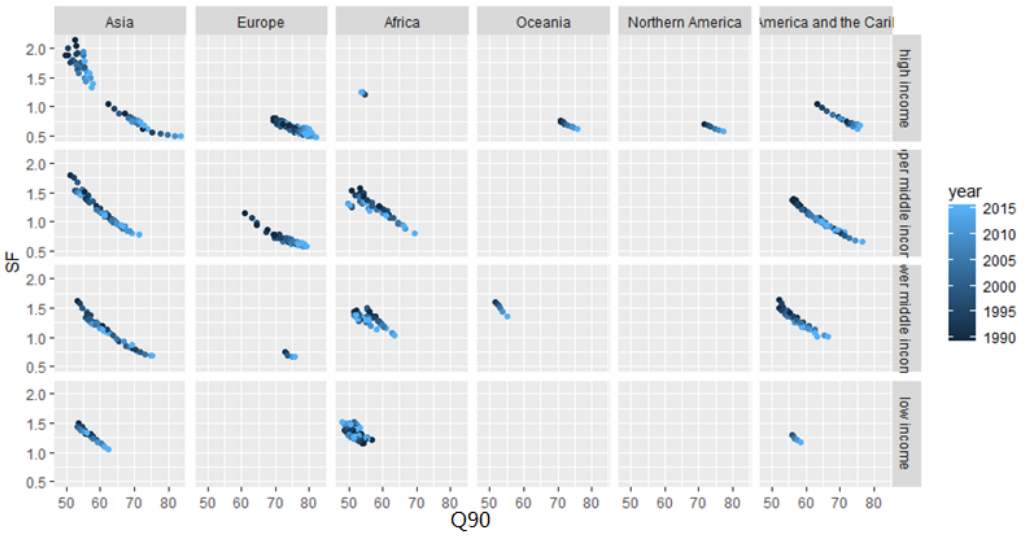
applied and the other one is more abstract. The applied one is as follows. There are many situations where someone may look at whether there are predictors of the mortality rate across countries. They might consider the following model:

$$log(\text{CMR}) = F(x_1, ..., x_n)$$

where x_i could be population-wise risk factors such as cigarette consumption, national income class, mental wellness, etc. The purpose is to see what explains $log(\text{CMR})$. Based on the motivation for age-standardization, it would be recognized that the above formula should take into account the influence of age structure. If people fail to consider the influence of age structure for this type of analysis, the analysis can still be partly rescued by putting Q90 into the formula. The abstract application is as follows. It is well-known that age is a classic confounder in the context of computing a single summary mortality rate (see for example our discussion about CMR at the beginning of Section 2). Our formula for SF gives a single-number adjustment for the confounder. Might there be other contexts where a similarly simple back of the envelop adjustment for confounding might be possible. This is an intriguing question which our analysis has raised. It is too early to tell but this will warrant future investigation.

While we focused on direct age-standardization, we speculate that our work could have insights for indirect method. Future work can explore whether SF, Q90, or other related constructs could have similar insights to offer with regard to indirectly age-standardized mortality rate.

Figure 22: SF *vs* Q90 for HMD and non-HMD countries, grouped by territory and income class, with WHO2000 as the reference population



9 Conclusion

ASMR has been widely used for comparing mortality risks across different populations. It is free from the confounding effect of the age structure, which is a key advantage over CMR, but at the cost of a more taxing requirement for data. While the calculation of CMR only requires the total population count and the total death count, the calculation of ASMR requires age-specific mortality rates, study population proportions and reference population proportions. Our model provides an alternative way of calculating ASMR even if age-specific mortality rates are unavailable. This model is linear and requires as input only the 90th quantiles of the study population distribution and the reference population, and CMR of the study population. To use our model, we suggest that users follow the following steps:

1. Calculate Q90 of the study population
2. If Q90 is greater than 60, proceed to step 3; Otherwise, stop using this model
3. Find reference Q90 and slope in Appendix (Table 17)
4. Plug into the model: $\text{SF} = 1 + \text{slope} \times (\text{Q90} - \text{Q90}_{\text{ref}})$

Age-standardization's effect on CMR has long been interpreted qualitatively. Standard epidemiology text books acknowledge that when the study population is younger than the reference population, then the ASMR of the study population will be greater than its CMR, vice versa. To our best knowledge, the existing literature stops at this qualitative understanding of age-standardization. The question is, how to define or quantify the 'oldness' of a population? What is the relationship between population oldness and age-standardization? Is this relationship linear, quadratic or exponential? Our analysis helps to answer these questions. According to our analysis, Q90 can be used as a reliable summary measure of population oldness, such that the relationship between age-standardization and population oldness, captured by Q90, is linear. For relatively young populations (Q90 < 60), such linearity is not as clear as it is for others. Finer data and further analysis are needed for more clarity.

Part II

Robust Risk Management under Covariance Uncertainty

Abstract

In the second part of the thesis, we consider the formulation of a general risk management procedure, where risk needs to be measured and further mitigated. The formulation admits an optimization representation and requires as input the distributional information about the underlying risk factors. Unfortunately, for most risk factors it is known to be difficult to identify their distribution in full details, and more problematically the risk management procedure can be prone to errors in the input distribution. In particular, one of the most important distribution information is the covariance that captures the spread and correlation among risk factors. We study the issue of covariance uncertainty in the problem of mitigating tail risk and by admitting an uncertainty set of covariance of risk factors, we propose a robust optimization model which minimizes risk for the worst scenario especially when data is insufficient and the number of risk factors is large. We will then transform our model into a computationally solvable one and test the model using real-world data.

10 Risk Management Based on Conditional Value-at-Risk (CVaR)

Risk management is a set of decisions made to reduce or prevent large losses. Let there be N stochastic risk factors $\boldsymbol{R} = (\mathcal{R}_1, ..., \mathcal{R}_N)^\top$ and $\boldsymbol{\pi} = (\pi_1, ..., \pi_N)^\top$ be the decision vector. Let L be a non-decreasing loss function, the loss in a system can be quantified by $L(\boldsymbol{\pi}, \boldsymbol{R})$. We want to make a decision $\boldsymbol{\pi}$ such that the corresponding loss distribution is the least risky. Since the true distribution of $\boldsymbol{R}$ is never known in practice and only a collection of T historical observations $\boldsymbol{R} = (\boldsymbol{R}_1, ..., \boldsymbol{R}_N)^\top$ are available, the empirical loss distribution $L(\boldsymbol{\pi}, \boldsymbol{R})$ is used. To quantify the risk of a loss distribution, a risk measure is needed.

Definition 1. Let $\mathcal{L}$ be a set of random variables. A function $\rho : \mathcal{L} \to \mathbb{R}$ is called a risk measure if it satisfies

(i) normalized: $\rho(0) = 0$,

(ii) monotone: $L_1,\ L_2 \in \mathcal{L},\ L_2 \geq L_1 \Rightarrow \rho(L_2) \geq \rho(L_1)$,

(iii) translative: $L \in \mathcal{L},\ a \in \mathbb{R} \Rightarrow \rho(L + a) = \rho(L) + a$.

Given a loss distribution $L(\boldsymbol{\pi}, \boldsymbol{R})$, the risk is measured as $\rho(L(\boldsymbol{\pi}, \boldsymbol{R}))$. An optimization problem seeks the optimal decision $\boldsymbol{\pi}$ that minimizes the risk captured by $\rho(L(\boldsymbol{\pi}, \boldsymbol{R}))$:

$$\min_{\boldsymbol{\pi}} \{\rho(L(\boldsymbol{\pi}, \boldsymbol{R}))\}$$

The appropriate choice of a risk measure ρ is dictated by the nature of uncertainties and risks in the problem at hand. Since risk managers are usually more concerned with large losses, i.e. right tail losses, downside risk measures that focus on the right tail of a loss distribution are reasonable choices, especially when the loss distribution is asymmetric and fat-tailed. In financial risk management, among the most popular downside risk measures are Value-at-Risk (VaR) and Conditional Value-at-Risk

(CVaR).

Given a loss distribution $L(\boldsymbol{\pi}, \boldsymbol{R})$, its cumulative probability function F_L and confidence level $\beta \in (0,1)$, β-level VaR is defined as

$$\text{VaR}_\beta(L(\boldsymbol{\pi}, \boldsymbol{R})) = \min\{l \mid F_L(l) \geq \beta\} = F_L^{-1}(\beta)$$

In other words, VaR_β is the β-quantile of L and it answers the question: what is the maximum loss with a specified confidence level β? For example, 90% VaR (i.e., 90th quantile) means that in 90% of cases the loss is expected to be smaller than the VaR amount, but it does not say anything about the size of losses in the rest 10% cases. VaR also has some undesirable mathematical characteristics such as a lack of subadditivity and convexity (Artzner et al., 1997,1999). Furthermore, VaR is difficult to optimize when it is calculated from scenarios. Mauser and Rosen (1999) and McKay and Keefer (1996) showed that VaR can be ill-behaved as a function of portfolio positions and can exhibit multiple local extrema.

As an alternative downside risk measure, Conditional Value-at-Risk (CVaR) has gained growing popularity in financial risk management. β-level CVaR can be roughly considered as the conditional expectation of the loss which exceeds the β-percentile of the loss distribution, i.e. the β-level VaR.

$$\text{CVaR}_\beta(L(\boldsymbol{\pi}, \boldsymbol{R})) = \int_\beta^1 \frac{1}{1-\beta} F_L^{-1}(p)dp$$

One of CVaR's advantages against VaR is that, CVaR takes into account extreme losses that are beyond the corresponding VaR threshold and thus ignored by VaR. CVaR is known to have better properties than VaR, such as convexity and coherency (a coherent risk measure satisfies sub-additivity and positive homogeneity in addition to the three basic properties of a risk measure, see Artzner et al., 1997; Embrechts, 1999; Pflug, 2000; Ogryczak and Ruszczynski, 2002). Additionally, Rockafellar and Uryasev (2000) show that the minimization of CVaR results in a convex optimization problem.

Given the aforementioned nice properties of CVaR, we will stick to CVaR as our

downside risk measure. Substitute ρ with CVaR in $\min_{\pi}\{\rho(L(\boldsymbol{\pi}, \boldsymbol{R}))\}$, we get the following β-level CVaR minimization model

$$\min_{\pi}\{\text{CVaR}_{\beta}(L(\boldsymbol{\pi}, \boldsymbol{R}))\}$$

CVaR minimization was first developed by Rockafellar and Uryasev (2002) and its numerical effectiveness was demonstrated in portfolio optimization and option hedging problems. Their work was then extended to objective functions consisting of different combinations of the expected loss and the CVaR, such as the minimization of the expected loss subject to a constraint on CVaR.

However, since CVaR is calculated by using only the tail distribution, a large number of samples are required for assuring statistical reliability. As a result of data shortage and data unreliability, tail distributions are often uncertain, and the CVaR minimization can be prone to error. For example, Takeda and Kanamori (2009) show that the estimation error in CVaR minimization is much more severe than that of the mean loss. Kondor et al. (2007) point out through experiments that CVaR suffers from more severe instability than variance and mean-absolute deviation. Recently, Lim et al. (2011) have demonstrated empirically the fragility associated with the CVaR minimization.

Our goal is to take into account the distribution of the uncertainty with respect to risk factor $\boldsymbol{R}$ and generate a robust CVaR minimization model that protects us against extreme losses.

11 Literature Review

11.1 Robust CVaR Minimization Model

In the previous section, we mentioned that CVaR minimization is exposed to estimation error of risk factor distribution. To alleviate such estimation error, robust optimization techniques have been developed. See Ben-Tal et al. (2009) for a comprehensive survey of recent developments in robust optimization. Robust optimizations take into account the least favorable scenario within a given set of possible scenarios, known as the uncertainty set. It should be noted that the structure of uncertainty set is critical in the sense that robust formulation with improper selection of the uncertainty set can result in a useless solution.

In particular, J. Gotoh et al. (2012) took into account the elliptical uncertainty set of the form

$$\mathcal{P} \coloneqq \left\{ \boldsymbol{\Delta R} \in \mathbb{R}^n \mid \boldsymbol{\Delta R}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\Delta R} \leq \delta^2 \right\}$$

where $\boldsymbol{\Sigma}$ could be any positive-definite and symmetric matrix (e.g., sample covariance) and $\delta > 0$ is a tuning parameter that controls the size of the uncertainty set.The following robust CVaR minimization model is constructed based on the uncertainty set $\mathcal{P}$

$$\min_{\pi} \max_{\boldsymbol{\Delta R} \in \mathcal{P}} \{\mathrm{CVaR}_\beta(L(\boldsymbol{\pi}, \boldsymbol{R} + \boldsymbol{\Delta R}))\}$$

They then proved that the above min-max model can be transformed into a regularized minimization model:

$$\min_{\pi} \{\mathrm{CVaR}_\beta(L(\boldsymbol{\pi}, L(\boldsymbol{\pi}, \boldsymbol{R}))) + \delta ||\boldsymbol{\pi}||^*\}$$

where $||\boldsymbol{\pi}||^* = \sqrt{\boldsymbol{\pi}^\top \boldsymbol{\Sigma} \boldsymbol{\pi}}$ is elliptical norm. The term $\delta ||\boldsymbol{\pi}||^*$ is known as regularization term which basically helps to alleviate over-fitting and to improve out-of-sample

performance (Vapnik, 2013). In particular, if $\mathbf{\Sigma}$ represents the sample covariance, then $\delta||\boldsymbol{\pi}||^*$ penalizes decisions that put heavy weights on assets with high volatility (variance).

The concept of robust optimization, i.e. min-max settings, is sometimes criticized for being overly conservative. However, this can be controlled by modifying the size of uncertainty set (i.e. tuning down δ).

11.2 A Robust Estimator for the Covariance Matrix

Covariance is one of the most important distribution information and is crucial to robust CVaR minimization model. In practice, the covariance needs to be estimated from sample, and as a result covariance uncertainty often exists due to the lack of observations (Jobson and Korkie, 1980). Thus, there is a need to take into account covariance uncertainty in order to make robust decisions. We will borrow the idea from robust estimator for covariance matrix to construct an uncertainty set for the covariance later on.

One methodology used to estimate covariance matrix is Bayesian method. Bayesian methods assume that we are computing an estimate of a distribution of covariance. This distribution is obtained by combining a "prior" with additional information. A number of Bayesian estimators have been proposed, for example Black-Litterman estimator (Stubbs and Vance, 2005 and Attilio Meucci, 2011).

Stein (1956) introduced the concept of shrinkage as an alternative to the ordinary least squares (OLS) estimator. In Bayesian analysis, shrinkage is defined in terms of priors. The philosophy of shrinkage is simple. Consider the sample covariance matrix $\boldsymbol{S}$ and a highly structured estimator (e.g., the identity matrix which is artificially structured and is independent of sample data), denoted by $\boldsymbol{F}$. We find a compromise between the two by computing a convex linear combination $\delta\boldsymbol{F} + (1-\delta)\boldsymbol{S}$, where δ is a number between 0 and 1. See Ledoit and Wolf (2003, 2004) for the general theory of shrinkage and applications to portfolio management, where they show that a shrinkage estimator yields significant improvements on actual stock return data.

Popular shrinkage estimators include Lasso estimator, Ridge estimator, and Stein-type estimators.

Notice that there is always a bias-variance trade-off in the shrinkage structure $\delta \boldsymbol{F} + (1-\delta)\boldsymbol{S}$. Such trade-off is controlled by the shrinkage constant δ. The higher the δ is, the more biased we are, vice versa. In one way or another, all successful risk models find a compromise between the sample covariance matrix $\boldsymbol{S}$ and the highly structured estimator $\boldsymbol{F}$ (e.g., identity matrix). Ledoit and Wolf (2003) proposed a formula for the optimal shrinkage constant δ which minimizes the expected distance between the shrinkage estimator and the true covariance matrix. In our work, we adopt a "shrinkage-like" estimator of the covariance matrix and choose the shrinkage constant under the robustness criteria, which we will elaborate more in later sections.

Other attempts to estimating the covariance matrix includes maximum likelihood (Roderick J. A. Little, 1988 and O. Banerjee et al., 2008) and principal component analysis (E. Candes et al., 2009 and A. D'Aspremont et al., 2008).

11.3 Robust Mean-Variance Model

When the distributions of risk factors are Gaussian, CVaR is proportional to the variance of the loss distribution (Bertsimas et al., 2004), and thus CVaR minimization is equivalent to mean-variance model (Markowitz, 1952). Mathematically, a mean-variance model is:

$$\min_{\boldsymbol{\pi}} \{var(L(\boldsymbol{\pi}, \boldsymbol{R}))\}$$

subject to

$$E(-L(\boldsymbol{\pi}, \boldsymbol{R})) \geq E_0$$

Typically, a mean-variance model seeks the decision $\boldsymbol{\pi}$ that minimizes the variance while keeping a minimum expected return, E_0. Here we ignore the expected return condition since this will not cause much loss according to Jagannanthan and Ma

(2003).

Like robust CVaR minimization, a robust mean-variance model can be constructed by taking into consideration an uncertainty set of model parameter(s) and admitting a min-max setting:

$$\min_{\boldsymbol{\pi}} \max_{p \in \mathcal{P}} \{var(L(\boldsymbol{\pi}, \boldsymbol{R}); p)\}$$

where p is any parameter that (partially) describes the underlying loss distribution, such as the distribution of risk factor $\boldsymbol{R}$. Several formulations of robust mean-variance model have been proposed. For example, Halldórsson and Tütüncü (2000) showed that if mean return vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ belong to the component-wise uncertainty sets $\mathcal{P}_{\boldsymbol{\mu}} = \left\{\boldsymbol{\mu} \mid \boldsymbol{\mu} \leq \boldsymbol{\mu} \leq \boldsymbol{\mu}^U\right\}$ and $\mathcal{P}_{\boldsymbol{\Sigma}} = \left\{\boldsymbol{\Sigma} \mid \boldsymbol{\Sigma} \leq \boldsymbol{\Sigma} \leq \boldsymbol{\Sigma}^U\right\}$, then the robust mean-variance model is equivalent to a non-linear saddle-point problem. Iyengar et al. (2003) argued that no procedure is provided for specifying the upper/lower bounds, $\left(\boldsymbol{\mu}^L, \boldsymbol{\mu}^U\right)$ and $\left(\boldsymbol{\Sigma}^L, \boldsymbol{\Sigma}^U\right)$, and that the solution algorithm is not practical when asset number is large. Ben-Tal et al. (2000) proposed another robust model where uncertainty sets are finite sets. Iyengar et al. (2003) considered a factor model for asset returns $\boldsymbol{R} = \boldsymbol{\mu} + \boldsymbol{V}^\top \boldsymbol{\mathcal{F}} + \boldsymbol{\epsilon}$, where $\boldsymbol{\mu}$ is the mean return vector, $\boldsymbol{\mathcal{F}}$ is the vector of factors that drives the market, $\boldsymbol{V}$ is the factor loading and $\boldsymbol{\epsilon}$ is the vector of residual returns. They further assumed independency between assets and considered an uncertainty set for covariance matrix $\boldsymbol{D}$[19] $\mathcal{P}_{\boldsymbol{D}} = \left\{\boldsymbol{D} \mid \boldsymbol{D} = \text{diag}(d),\ d_i \in \left[\underline{d_i}, \overline{d_i}\right]\right\}$, an elliptical uncertainty set for $\boldsymbol{V}$, and a box-type uncertainty set for $\boldsymbol{\mu}$. The robust mean-variance model becomes a second order conic problem which can be efficiently solved.

The following sections are organized as follow. In Section 12, we briefly review mean-variance model under normality assumption and introduce robust covariance method. In Section 13, we consider CVaR minimization without normality assumption and review a robust CVaR minimization model proposed by J. Gotoh et al. (2012). By arguing that their model is incapable of capturing information of large

[19]Notation D is used because the covariance matrix is diagonal in this case.

losses, we suggest a revised version which is robust in the sense that 1) it incorporates both covariance uncertainty and distribution uncertainty of the risk factors; 2) it captures two types of risk, namely occasional large losses and overall high volatility. In Section 14, we implement our model in the context of portfolio management and justify its advantages by applying on real-world data.

12 Robust Mean-Variance Model

In this section, we will base our discussion on normality assumption. We will introduce a mean-variance model in subsection 12.1. In subsection 12.2, we will construct an uncertainty set for the covariance matrix and develop a robust mean-variance model using a min-max approach.

12.1 Mean-Variance Optimization

We introduce our notations in the context of portfolio management.

Suppose a portfolio consists of N assets, whose returns are random variables $\boldsymbol{R} = (\mathcal{R}_1, ..., \mathcal{R}_N)^\top$. Let $\boldsymbol{R} = (\boldsymbol{R}_1, ..., \boldsymbol{R}_N)^\top$ be a collection of T historical observations of $\boldsymbol{R}$. Let $\boldsymbol{\pi} = (\pi_1, ..., \pi_N)^\top$ denote the proportion of wealth invested in asset $1, ..., N$. The total return of the portfolio is $\sum_{n=1}^{N} \pi_n \boldsymbol{R}_n = \boldsymbol{R}^\top \boldsymbol{\pi}$. Let the loss function simply be $L(\boldsymbol{\pi}, \boldsymbol{R}) = -\boldsymbol{R}^\top \boldsymbol{\pi}$. Let the risk be measured by the variance of the loss function $var(L(\boldsymbol{\pi}, \boldsymbol{R})) = \boldsymbol{\pi}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{\pi}$, where $\hat{\boldsymbol{\Sigma}} = cov(\boldsymbol{R})$ denotes the sample covariance. A portfolio optimization problem seeks to find a decision $\boldsymbol{\pi}$ that minimizes the portfolio risk captured by $\boldsymbol{\pi}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{\pi}$:

$$\min_{\boldsymbol{\pi}} \left\{ \boldsymbol{\pi}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{\pi} \right\}$$

$$\boldsymbol{\pi} \in \left\{ \boldsymbol{\pi} \mid \sum_{n=1}^{N} \pi_n = 1,\ \pi_n \geq 0 \right\}$$

Note that the constraint $\sum_{n=1}^{N} \pi_n = 1$ indicates that the total wealth is normalized to 1. The non-negativity constraint $\pi_n \geq 0$ can be interpreted as "no-short-sale" constraint (here we follow Jagannanthan and Ma (2003) and include the non-negativity constraint). Hereafter these two constraints will be taken into account by default and will not be displayed for simplicity.

12.2 Robust Mean-Variance Optimization under Covariance Uncertainty

The aforementioned optimization model $\min_{\boldsymbol{\pi}} \left\{ \boldsymbol{\pi}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{\pi} \right\}$ is based on the sample covariance matrix. It has long been recognized that the resulting solutions perform poorly out-of-sample, because sample covariance matrix is an imprecise estimator of covariance matrix, especially when the dimension is high. Here we propose a spectrum of covariance matrices $\mathcal{Z}$, in order to capture the covariance uncertainty. A robust decision is then based on this covariance uncertainty set. We look for the solution $\boldsymbol{\pi}$ that minimizes $\boldsymbol{\pi}^\top \boldsymbol{\Sigma} \boldsymbol{\pi}$ when the worst realization of covariance in $\mathcal{Z}$ happens. Mathematically, we formulate the following robust min-max model:

$$\min_{\boldsymbol{\pi}} \max_{\boldsymbol{\Sigma} \in \mathcal{Z}} \left\{ \boldsymbol{\pi}^\top \boldsymbol{\Sigma} \boldsymbol{\pi} \right\} \tag{1}$$

Our goal in this section is to construct the covariance uncertainty set $\mathcal{Z}$.

Consider the following principal component decomposition of sample covariance:

$$\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}}_n + \hat{\boldsymbol{\Sigma}}_{N-n}$$

where $\hat{\boldsymbol{\Sigma}}_n$ consists of the first n principal components and $\hat{\boldsymbol{\Sigma}}_{N-n}$ consists of the rest $N-n$ principal components (see more details in appendix). We regard $\hat{\boldsymbol{\Sigma}}_n$ as the main source of information and de-noise $\hat{\boldsymbol{\Sigma}}_{N-n}$ by diagonalization. The covariance matrix is estimated as:

$$\hat{\boldsymbol{\Sigma}}(n) = \hat{\boldsymbol{\Sigma}}_n + diag(\hat{\boldsymbol{\Sigma}}_{N-n})$$

Notice that as the constant n decreases from N to 1, the estimator $\hat{\boldsymbol{\Sigma}}(n)$ shrinks from $\hat{\boldsymbol{\Sigma}}(N)$ (sample covariance) to $\hat{\boldsymbol{\Sigma}}(1)$. The sample covariance $\hat{\boldsymbol{\Sigma}}(N)$ has the property of being unbiased (i.e., its expected value is equal to the true covariance matrix). However it contains a lot of estimation error when the number of data points is of comparable or even smaller order than the number of risk factors. $\hat{\boldsymbol{\Sigma}}(1)$, on the other

hand, has a lot of structure (e.g., diagonalization). It contains relatively little estimation error but can be biased. To take into account as many possibilities between the sample covariance matrix and the highly structured $\hat{\mathbf{\Sigma}}(1)$, we construct the covariance uncertainty set as a linear combination of $\left\{\hat{\mathbf{\Sigma}}(n)\right\}$:

$$\mathcal{Z} = \left\{ \sum_{n=1}^{N} w_n \hat{\mathbf{\Sigma}}(n) \mid \sum_{n=1}^{N} w_n = 1,\, w_n \geq 0 \right\}$$

In this way, we capture different levels of variance and bias.

As a supplementary detail, the following fact shows that diagonalization helps to reduce the skewness of any semi-definite symmetric matrix, which is why we use diagonalization to de-noise $\hat{\mathbf{\Sigma}}_{N-n}$.

Fact 2. For any semi-definite symmetric matrix $\mathbf{\Sigma}$, its conditional number, defined as the largest eigenvalue divided by the smallest eigenvalue, will always be smaller (if not unchanged) after diagonalization.

Proof. (See appendix.)

□

Finally, we apply duality to transform problem (1) into a minimization problem which can be efficiently programmed and solved.

By definition of $\mathcal{Z}$, $\max_{\mathbf{\Sigma} \in \mathcal{Z}} \left\{ \boldsymbol{\pi}^\top \mathbf{\Sigma} \boldsymbol{\pi} \right\}$ is equivalent to

$$\max_{\boldsymbol{w}} \left\{ [\boldsymbol{\pi}^\top \hat{\mathbf{\Sigma}}(1) \boldsymbol{\pi}, ..., \boldsymbol{\pi}^\top \hat{\mathbf{\Sigma}}(N) \boldsymbol{\pi}]^\top \boldsymbol{w} \right\}$$

where

$$\sum_{n=1}^{N} w_n = 1,\, w_n \geq 0$$

According to strong[20] linear duality, the above problem is equivalent to

[20]Strong duality requires that there is a feasible solution, which is obvious in this case.

$$\min_{v} \{v\}$$

where

$$v \geq \boldsymbol{\pi}^{\top} \hat{\boldsymbol{\Sigma}}(n) \boldsymbol{\pi},\ n = 1, ..., N$$

Thus, $\min_{\boldsymbol{\pi}} \max_{\boldsymbol{\Sigma} \in \mathcal{Z}} \left\{ \boldsymbol{\pi}^{\top} \boldsymbol{\Sigma} \boldsymbol{\pi} \right\}$ is equivalent to

$$\min_{v, \boldsymbol{\pi}} \{v\}$$

subject to

$$v \geq \boldsymbol{\pi}^{\top} \hat{\boldsymbol{\Sigma}}(n) \boldsymbol{\pi},\ \sum_{n=1}^{N} \pi_n = 1,\ \pi_n \geq 0,\ n = 1, ..., N$$

The constraints suggest that this is a corner point problem and that $\boldsymbol{\Sigma}(worst) \in \left\{ \hat{\boldsymbol{\Sigma}}(1), ..., \hat{\boldsymbol{\Sigma}}(N) \right\}$.

13 Robust CVaR Minimization Model

In the previous section, we made a normality assumption and discussed a robust mean-variance model under covariance uncertainty. In this section we remove normality assumption and consider the CVaR minimization under distribution uncertainty. We follow Gotoh et al. (2012) where they proposed a robust CVaR minimization model. Our contribution is to bring covariance uncertainty into consideration and propose a even more robust CVaR minimization model.

13.1 Robust Factor-Based CVaR Minimization under Distribution Uncertainty

In this subsection, we briefly review a robust CVaR minimization model proposed in Gotoh et al. (2012).

Recall that a general CVaR minimization model is

$$\min_{\boldsymbol{\pi}} \{\text{CVaR}_\beta(L(\boldsymbol{\pi}, \boldsymbol{R}))\}$$

Gotoh et al. (2012) argued that a perturbation in large loss scenarios can make a big impact on the calculation of CVaR. In order to alleviate the effect of such a perturbation, they followed Konno et al. (2002) and replaced the observed risk factor $\boldsymbol{R}$ with values estimated by a factor approach $\boldsymbol{R} = \boldsymbol{\mathcal{R}}_F + \epsilon = \underbrace{\boldsymbol{\mu} + \boldsymbol{V}^\top \boldsymbol{F}}_{\boldsymbol{R}_F} + \boldsymbol{\epsilon}$, where $\boldsymbol{\mu} \in \mathbb{R}^N$ is the vector of mean returns, $\boldsymbol{F} \in \mathbb{R}^M$ is the vector of M factors that drives the market[21], $\boldsymbol{V} \in \mathbb{R}^{M \times N}$ is the factor loading matrix and $\boldsymbol{\epsilon}$ is the vector of residual returns. They then considered the following distribution uncertainty for risk factor $\boldsymbol{R}$

$$\mathcal{P}_1 = \left\{ \boldsymbol{\Delta R} \in \mathbb{R}^N \mid \boldsymbol{\Delta R}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\Delta R} \leq \delta^2, \boldsymbol{\Sigma} = cov(\boldsymbol{R} - \boldsymbol{R}_F) \right\}$$

The resulting robust factor-based CVaR minimization model is therefore

[21] Gotoh et al. (2012) used Fama-French factors for $\boldsymbol{F}$

$$\min_{\boldsymbol{\pi}} \max_{\boldsymbol{\Delta R} \in \mathcal{P}_1} \{\text{CVaR}_\beta(L(\boldsymbol{\pi}, \boldsymbol{R}_F + \boldsymbol{\Delta R}))\} \tag{2}$$

It is robust in the sense that it looks for the solution $\boldsymbol{\pi}$ which minimizes β-level CVaR when the worst realization of distribution uncertainty in $\mathcal{P}_1$ happens.

13.2 Robust Non-Factor CVaR Minimization under Distribution Uncertainty

We argue that the above factor-based CVaR minimization model has a problem of information loss. Our goal in this subsection is to revise the model in order to retain the lost information and also take into account covariance uncertainty.

According to Corollary 3.2 in Gotoh et al. (2012) and Rockafellar and Uryasev (2000), the objective function in (2) is equivalent to

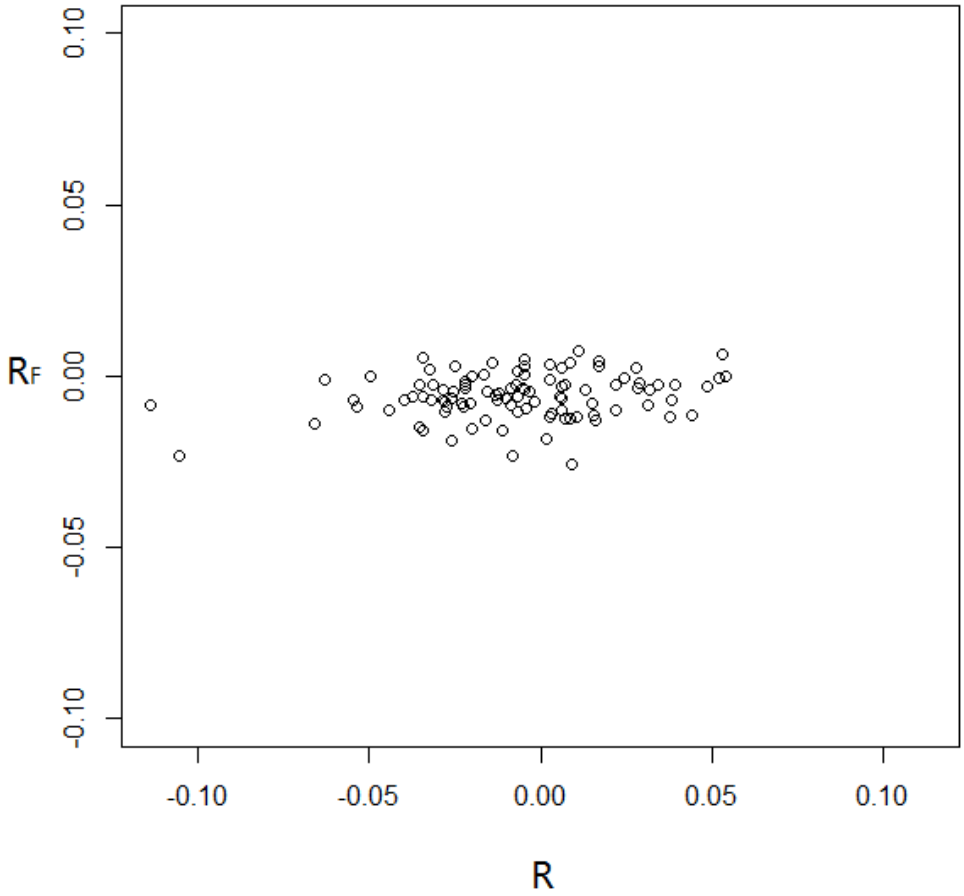
$$\text{CVaR}_\beta(L(\boldsymbol{\pi}, \boldsymbol{R}_F)) + \delta\sqrt{var(L(\boldsymbol{\pi}, \boldsymbol{R} - \boldsymbol{R}_F))} \tag{3}$$

which contains two parts: a factor-model based CVaR, $\text{CVaR}_\beta(L(\boldsymbol{\pi}, \boldsymbol{R}_F))$, and a scaled standard deviation of the residual loss, $\delta\sqrt{var(L(\boldsymbol{\pi}, \boldsymbol{R} - \boldsymbol{R}_F))}$. The first part is responsible for capturing large losses, while the second part is responsible for capturing overall volatility.

However, we find empirically that the goodness-of-fit of the factor model $\boldsymbol{R} = \boldsymbol{\mu} + \boldsymbol{V}^\top \boldsymbol{F} + \boldsymbol{\epsilon}$ is low[22], see Figure 23. As a consequence, $L(\boldsymbol{\pi}, \boldsymbol{R}_F)$ poorly estimates $L(\boldsymbol{\pi}, \boldsymbol{R})$, and thus the factor model-based CVaR loses information about large losses. Moreover, this information loss cannot be retained by regularization term. This is because the regularization term is a geometric mean which would lower the impact of a few large losses. In fact, when the factor model has low predictive power, in other words $\boldsymbol{R}_F = \boldsymbol{R} - \boldsymbol{\epsilon} \simeq 0$, $\text{CVaR}_\beta(L(\boldsymbol{\pi}, \boldsymbol{R}_F))$ is almost equal to constant 0, and the factor-based CVaR minimization is essentially the same as mean-variance

[22] We have an adjusted R-square of 0.009373, an F-statistics of 1.312, and a p-value of 0.2749. All these statistics suggest that the factor model has a low goodness-of-fit.

Figure 23: Goodness-of-fit of factor model



optimization.

To fix the problem of information loss, we modify (2) as

$$\min_{\boldsymbol{\pi}} \max_{\boldsymbol{\Delta R}\in\mathcal{P}_2} \{\text{CVaR}_\beta(L(\boldsymbol{\pi}, \boldsymbol{R}+\boldsymbol{\Delta R}))\} \tag{4}$$

where

$$\mathcal{P}_2 = \left\{\boldsymbol{\Delta R} \in \mathbb{R}^N \mid \boldsymbol{\Delta R}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\Delta R} \leq \delta^2,\, \boldsymbol{\Sigma} = cov(\boldsymbol{R})\right\}$$

Substituting $\boldsymbol{R}_F$ with $\boldsymbol{R}$ and $\mathcal{P}_1$ with $\mathcal{P}_2$, we again follow Corollary 3.2 in Gotoh et al. (2012) to transform the objective function in (4) into

$$\text{CVaR}_\beta(L(\boldsymbol{\pi}, \boldsymbol{R})) + \delta\sqrt{var(L(\boldsymbol{\pi}, \boldsymbol{R}))} \tag{5}$$

The difference between (3) and (5) lies in the CVaR component. When the factor model has low goodness-of-fit, $\boldsymbol{R} \simeq \boldsymbol{\epsilon}$. In this case, $\text{CVaR}_\beta(L(\boldsymbol{\pi}, \boldsymbol{R}))$ in (5) is the 'empirical' CVaR and is more meaningful and informative than its counterpart in (3). Since $\boldsymbol{R} \simeq \boldsymbol{\epsilon}$, the regularization part of (3) and (5) makes no big difference.

We refer to this model as non-factor CVaR minimization. Our non-factor model is better at capturing information for extreme losses, since we admit the empirical CVaR. Meanwhile, the regularization term takes care of high overall volatility. Our model also has a good asymptotic property. As tuning parameter δ approaches infinity, our model becomes mean-variance model that we discussed at the beginning of this paper; as δ approaches 0, our model becomes CVaR minimization model. However, the factor based-model does not have this property.

13.3 Robust Non-Factor CVaR Minimization under Distribution and Covariance Uncertainty

So far, the model is robust in the sense that distribution uncertainty $\mathcal{P}_2$ has been incorporated. Notice that $\mathcal{P}_2$ is built upon covariance which is estimated by means

of the sample covariance. In order to mitigate the estimation error of covariance, we further robustify the model by taking into account covariance uncertainty $\mathcal{Z}$ and look for the solution $\boldsymbol{\pi}$ that minimizes β-CVaR when the worst realization of covariance in $\mathcal{Z}$ happens. Mathematically, we formulate the following min-max-max optimization problem:

$$\min_{\boldsymbol{\pi}} \max_{\boldsymbol{\Sigma}\in\mathcal{Z}} \max_{\boldsymbol{\Delta R}\in\mathcal{P}_2} \{\mathrm{CVaR}_\beta(L(\boldsymbol{\pi}, \boldsymbol{R}+\boldsymbol{\Delta R}))\} \tag{6}$$

We apply the same covariance uncertainty constructed in subsection 12.2:

$$\mathcal{Z} = \left\{ \sum_{n=1}^{N} w_n \hat{\boldsymbol{\Sigma}}(n) \mid \sum_{n=1}^{N} w_n = 1,\ w_n \geq 0 \right\}$$

Lemma 3. *Optimization problem (6) with a linear loss function* $L(\boldsymbol{\pi}, \boldsymbol{R}+\boldsymbol{\Delta R}) = (\boldsymbol{R}+\boldsymbol{\Delta R})^\top \boldsymbol{\pi}$ *is equivalent to*

$$\min_{\boldsymbol{\pi},\alpha,v,\boldsymbol{y}} \left\{ \alpha + \frac{1}{(1-\beta)T} \boldsymbol{e}^\top \boldsymbol{y} + \delta u \right\} \tag{7}$$

subject to

$$\boldsymbol{y} \geq -\boldsymbol{R}^\top \boldsymbol{\pi} - \alpha,\ \boldsymbol{y} \geq 0$$

$$u \geq \sqrt{\boldsymbol{\pi}^\top \boldsymbol{\Sigma}(n) \boldsymbol{\pi}},\ n = 1, 2, ..., N$$

$$\boldsymbol{e} = \underbrace{(1, ..., 1)}_{T\times 1}^\top$$

(T is the number of observations.)

Proof. According to corollary 3.2 (Gotoh et al., 2012) problem (6) with a linear loss function $L(\boldsymbol{\pi}, \boldsymbol{R}+\boldsymbol{\Delta R}) = (\boldsymbol{R}+\boldsymbol{\Delta R})^\top \boldsymbol{\pi}$ is equivalent to

$$\min_{\alpha,\boldsymbol{\pi},\boldsymbol{y}} \max_{\boldsymbol{w}} \left\{ \alpha + \frac{1}{(1-\beta)\,T} \boldsymbol{e}^\top \boldsymbol{y} + \delta \sqrt{\boldsymbol{\pi}^\top \left(\sum_{n=1}^{N} w_n \hat{\boldsymbol{\Sigma}}(n) \right) \boldsymbol{\pi}} \right\}$$

where $\boldsymbol{y} \geq -\boldsymbol{R}^\top \boldsymbol{\pi} - \alpha$, $\boldsymbol{y} \geq 0$, and T denotes the number of observations.

$$= \min_{\alpha,\boldsymbol{\pi},\boldsymbol{y}} \max_{\boldsymbol{w}} \left\{ \alpha + \frac{1}{(1-\beta)\,T} \boldsymbol{e}^\top \boldsymbol{y} + \delta \sqrt{\sum_{n=1}^{N} w_n \left(\boldsymbol{\pi}^\top \hat{\boldsymbol{\Sigma}}(n) \boldsymbol{\pi} \right)} \right\}$$

$$= \min_{\alpha,\boldsymbol{\pi},\boldsymbol{y}} \left\{ \alpha + \frac{1}{(1-\beta)\,T} \boldsymbol{e}^\top \boldsymbol{y} + \delta \sqrt{\max_{\boldsymbol{w}} \sum_{n=1}^{N} w_n \left(\boldsymbol{\pi}^\top \hat{\boldsymbol{\Sigma}}(n) \boldsymbol{\pi} \right)} \right\}$$

We already proved that $\max_{\boldsymbol{w}} \sum_{n=1}^{N} w_n \left(\boldsymbol{\pi}^\top \hat{\boldsymbol{\Sigma}}(n) \boldsymbol{\pi} \right)$ is equivalent to $\min_{v} \{v\}$, where $v \geq \boldsymbol{\pi}^\top \hat{\boldsymbol{\Sigma}}(n) \boldsymbol{\pi}$, $n = 1, ..., N$. Therefore, optimization problem (6) with a linear loss function $L(\boldsymbol{\pi}, \boldsymbol{R} + \boldsymbol{\Delta R}) = (\boldsymbol{R} + \boldsymbol{\Delta R})^\top \boldsymbol{\pi}$ is equivalent to (7):

$$\min_{\boldsymbol{\pi},\alpha} \left\{ \alpha + \frac{1}{(1-\beta)T} \boldsymbol{e}^\top \boldsymbol{y} + \delta u \right\}$$

subject to

$$\boldsymbol{y} \geq -\boldsymbol{R}^\top \boldsymbol{\pi} - \alpha, \boldsymbol{y} \geq 0$$

$$u \geq \sqrt{\boldsymbol{\pi}^\top \boldsymbol{\Sigma}(n) \boldsymbol{\pi}},\ n = 1, 2, ..., N$$

□

This dual transformation is useful because problem (7) can be tractably solved as a second order conic problem.

14 Numerical Experiments

In this section, we will demonstrate by numerical experiments the usefulness of two proposed methods, namely robust covariance and non-factor CVaR minimization. Recall that robust covariance method is born for data insufficiency. We will show that when sample size is small (relative to the dimension of problem), we benefit from robust covariance method. On the other hand, when sample size is relatively large, we benefit from non-factor model while factor-model suffers from in-sample information loss.

When assessing the performance of the two proposed methods, our criteria is robustness and our primary performance measure is out-of-sample CVaR, which aligns with our intention to control extreme out-of-sample loss. Since one single CVaR is insufficient and lacks statistical power, we run the model 100 times, which gives us 100 CVaRs. Then we calculate the *average* of the *worst five* CVaRs. A robust model should provide good protection under worst-case scenarios, and as a result, the *average* should be smaller than that of a non-robust model. We further define an *excess ratio* to facilitate comparison of *average* between non-robust and robust models. The *excess ratio* is calculated as the percentage of decrease in the *average.* For example, if the *average* is 100 for the non-robust model and 90 for the robust model, then the *excess ratio* will be $\frac{100-90}{100} \times 100\% = 10\%$. A positive *excess ratio* like this means that robust model is better than non-robust one, and vice versa.

Our data source is *Yahoo Finance,* where we grabbed 10-year daily stock price of 120 stocks (around 2700 observations in total). From that, we constructed 100 subsets on a rolling horizon scheme. Each subset is further divided into a training set (sample) and a testing set (out-of-sample). Note that as sample size becomes large, those 100 subsets may overlap on each other. The data experiments were coded and run in *Matlab 2016a.*

Here we introduce our parameter settings. We did not mean to exhaust all possible parameter settings and give all the explanation but focused on cases where there is

more regularity.

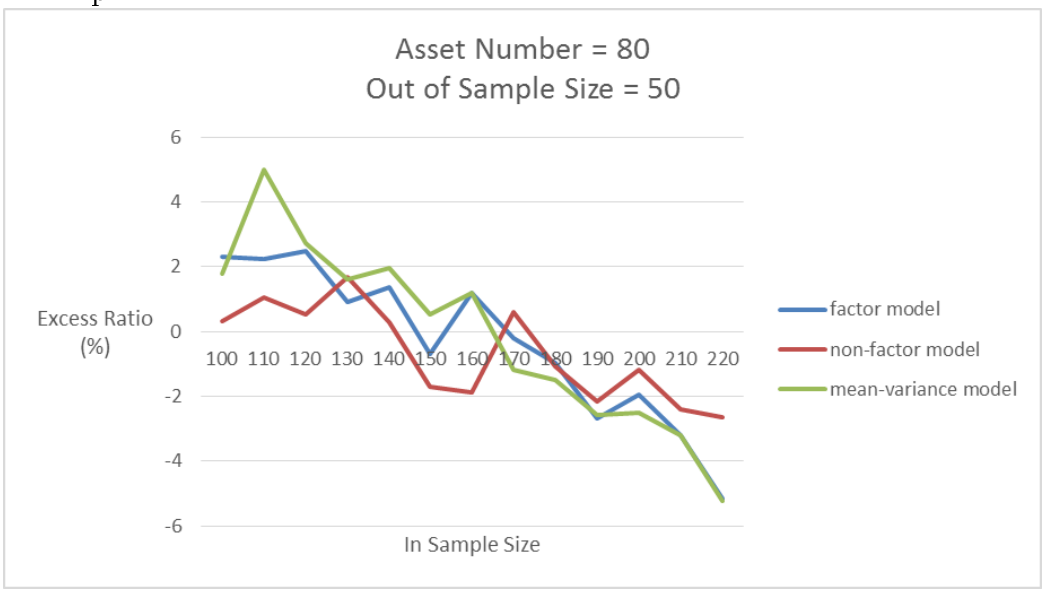
- We limit our discussion to a fixed out-of-sample size of 50. A future period of 50 days is fairly long enough to test model performance. From a technical point of view, increasing out-of-sample size distorts out-of-sample CVaR, leading to an unfair comparison of model performance.
- We choose asset number 80 to be a benchmark case. This is representative of a high dimensional case.
- $\beta = 0.99$ and $\delta = 10$. Note that with an out-of-sample size of 50, the 0.99-level CVaR equals to the maximum loss. Remember that δ is the scaling parameter and $\delta = 10$ stands for a hybrid of mean-variance model and CVaR minimization model.

14.1 Performance Comparisons between Robust Covariance and Sample Covariance

In this subsection, we will compare the performance of robust covariance matrix against that of sample covariance matrix in three different models, namely mean-variance model, factor-based CVaR minimization model, and non-factor CVaR minimization model. For each one of these three models, we separately applied sample covariance and robust covariance method and calculated an *excess ratio*. For example, in Figure 24 we see that when sample size is 100 and mean-variance model is used, the *excess ratio* is around 2%. This means that robust covariance method is more robust than sample covariance when it is applied for mean-variance model and sample size is 100.

We can also observe that for all three models, the *excess ratio* is generally better when sample size is relatively small. This indicates that when asset number is 80 and data is relatively insufficient, our robust covariance method performs better than

Figure 24: Performance test for robust covariance approach, average of the worst 5 out-of-sample CVaRs



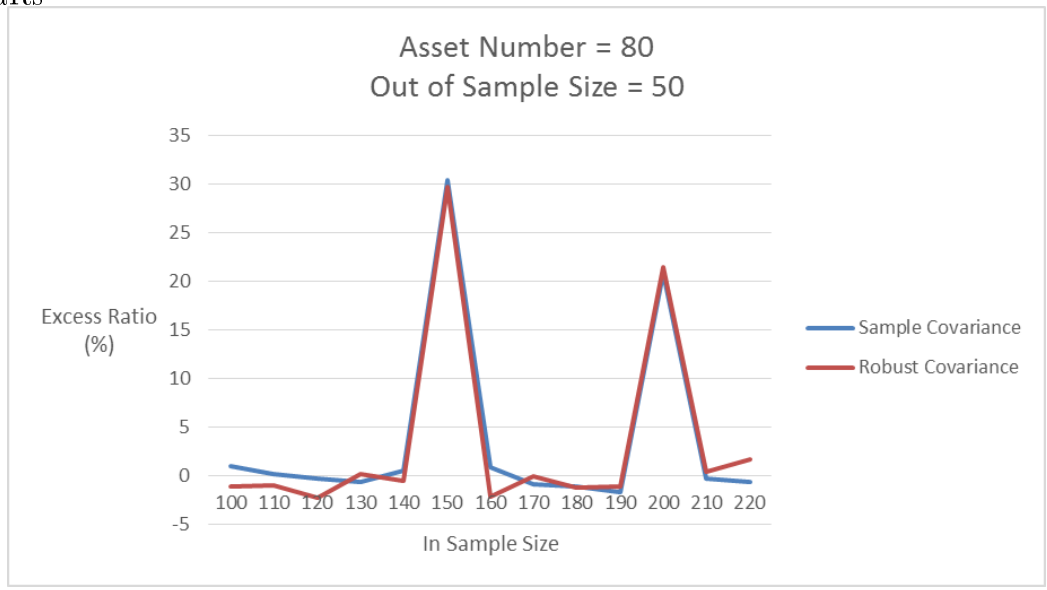
sample covariance. Notice that *excess ratio*s are negative when sample size is relatively large. This is because when data is relatively sufficient, sample covariance is a better estimate of population covariance.

14.2 Performance Comparisons between Factor-Based and Non-Factor Model

In this subsection, we will compare the performance of non-factor CVaR model against that of factor-based CVaR model. Firstly, we used sample covariance for both two models and calculated an *excess ratio*. Then we switched to robust covariance method and repeated the above process. For example, in Figure 25, we see that when sample size is 100 and sample covariance is used, the *excess ratio* is around 0, which means that in that case, factor-based and non-factor model have similar performance.

However, there are cases when the advantage of non-factor model is apparent. For example, when sample size is 150, the *excess ratio* is around 30%; when sample size is 200, the *excess ratio* is around 20%. This is because CVaR model works well for these two cases and non-factor model benefits from that since it is a combination of CVaR model and mean-variance model. In contrast, factor-based model is essentially

Figure 25: Performance test for non-factor model, average of the worst 5 out-of-sample CVaRs



the same as mean-variance and takes no advantage of CVaR model.

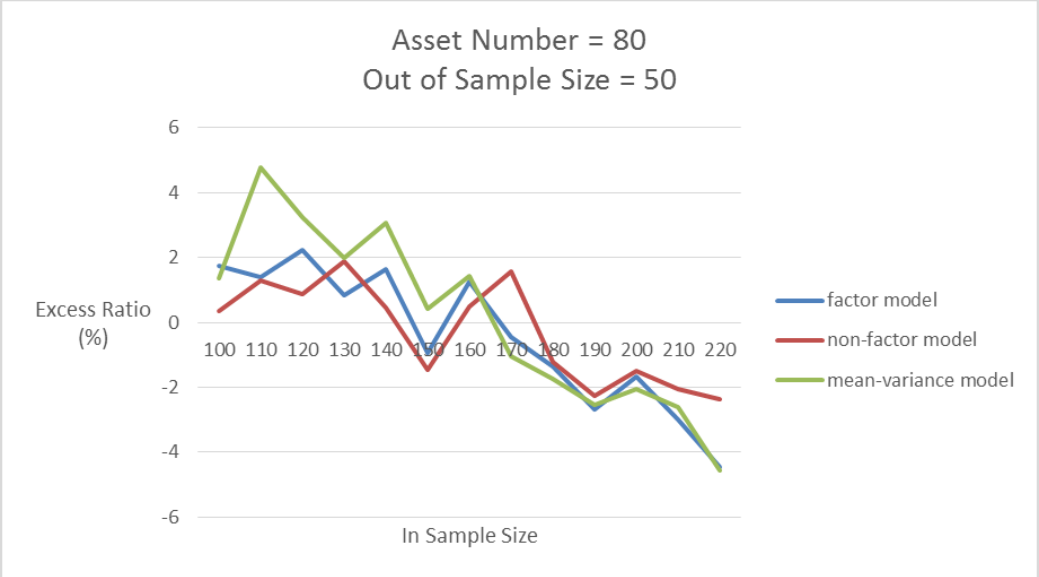
Notice that according to our existing numerical experiments, *excess ratio* is no worse than -5%, while it can be as good as 30%. It is clear that in our cases, implementing non-factor model is beneficial on average.

14.3 Discussion

Our main experimental results, as shown in Figure 24, align with our methodology. We see that as sample size becomes smaller, our model generally provides better protection against extreme risks. Meanwhile, we notice that the trend (the smaller the sample size, the better our model's performance) is not perfectly monotone. This is because real data comes with randomness, and when sample size is relatively small, our covariance uncertainty set cannot capture every possible realization of covariance.

We chose the average of worst 5 out-of-sample CVaR as our performance measure. Obviously, different percentages can be used instead of 5%. Sensitivity analysis helps to examine the impact of different percentages employed. For example, Figure 26 shows model performance if 6% is used (i.e. the average of worst 6 out-of-sample CVaR). We find insignificant change in general trends. One might think of using the worst case CVaR as the performance measure, given that our methodology is

Figure 26: Performance test for robust covariance approach, average of the worst 6 out-of-sample CVaRs



designed to provide protection under the worst case scenario. However, we argue that the worst case CVaR is statistically unstable because of the randomness of the data. In comparison, it is statistically more stable to use the average of the higher quantiles as a performance measure.

15 Conclusion

For the second part of the thesis, our main contribution is that we proposed a new method to take into account the uncertainty of covariance. By considering covariance uncertainty, we construct a worst case optimization model which finds the optimal solution against the least favourable covariance. To construct the uncertainty set of covariance, we used principal component analysis and diagonalization approach. The numerical results generally align with our goal which is to provide better protection against extreme risks when sample size is relatively small. Our second contribution is to provide a non-factor CVaR minimization model which is supposed to take more advantage of sample data when sample size is reasonably large. Future work needs to tell whether the sample size is large enough or small enough. To take advantage of our proposed models, we also suggest the following two combinations: 1) to use sample covariance with non-factor model when sample size is relatively large and 2) to consider covariance uncertainty and use factor-based model when sample size is relatively small. Ideally, future work will establish a criteria to recommend which combination to be used.

References

[1] Adjuik, M., Smith, T., Clark, S., Todd, J., Garrib, A., Kinfu, Y., ... & Adazu, U., 2006. Cause-specific mortality rates in sub-Saharan Africa and Bangladesh. Bulletin of the World Health Organization, 84(3), 181-188.

[2] Anderson, R. N., & Rosenberg, H. M., 1998. Age standardization of death rates: implementation of the year 2000 standard. National vital statistics reports, 47(3), 1-17.

[3] Belsley, D.A., Kuh, E. and Welsch, R.E., 1980. Regression diagnostics: Identifying influential data and sources of collinearity.

[4] Bertsimas, D., Brown, D. B., & Caramanis, C., 2011. Theory and applications of robust optimization. SIAM review, 53(3), 464-501.

[5] Bertsimas, D., Lauprete, G. J., & Samarov, A., 2004. Shortfall as a risk measure: properties, optimization and applications. Journal of Economic Dynamics and control, 28(7), 1353-1381.

[6] Brownlee, J., & Young, M., 1922. The epidemiology of summer diarrhoea.

[7] Butenko, S., Murphey, R., & Pardalos, P. M. (Eds.)., 2013. Cooperative control: models, applications and algorithms (Vol. 1). Springer Science & Business Media.

[8] Carlo, A., & Prospero, S., 2002. Portfolio optimization with spectral measures of risk. arXiv preprint cond-mat/0203607.

[9] Chen, D. H., Chen, C. D., & Chen, J., 2009. Downside risk measures and equity returns in the NYSE. Applied Economics, 41(8), 1055-1070.

[10] Chen, Y., Wiesel, A., & Hero, A. O., 2011. Robust shrinkage estimation of high-dimensional covariance matrices. IEEE Transactions on Signal Processing, 59(9), 4097-4107.

[11] Chiang, C. L., 1961 Standard error of the age-adjusted death rate.

[12] Chiang, C.L. and World Health Organization, 1979. Life table and mortality analysis.

[13] Chow, Y., Tamar, A., Mannor, S., & Pavone, M., 2015. Risk-sensitive and robust decision-making: a cvar optimization approach. In Advances in Neural Information Processing Systems (pp. 1522-1530).

[14] Curtin, L. R., & Klein, R. J., 1995. Direct standardization (age-adjusted death rates) (No. 6). US Department of Health and Human Services, Public Health Service, Centers for Disease Control and Prevention, National Center for Health Statistics.

[15] Dempsev, M., 1947. Decline in tuberculosis; the death rate fails to tell the entire story. American Review of Tuberculosis 86:157.

[16] Dickinson, F G and Welker, E L, 1948. What is the leading cause of death? Two new measures. Bureau of Medical Economic Research, American Medical Association, Bulletin 64, Chicago.

[17] Draper, N.R. and Smith, H., 1998. Fitting a straight line by least squares. Applied regression analysis, pp.15-46.

[18] Evgeniou, T., Pontil, M., & Poggio, T., 2000. Regularization networks and support vector machines. Advances in computational mathematics, 13(1), 1.

[19] Fabozzi, F. J., Kolm, P. N., Pachamanova, D. A., & Focardi, S. M., 2007. Robust portfolio optimization and management. John Wiley & Sons.

[20] General Register Office, 1841,1853,1857,1883, 1884. Annual Report of the Registrar General for England and Wales.

[21] Gillum, R. F., 2002. New considerations in analyzing stroke and heart disease mortality trends: the Year 2000 Age Standard and the International Statistical Classification of Diseases and Related Health Problems, 10th Revision. Stroke, 33(6), 1717-1722.

[22] Global Health Estimates 2015: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2015. Geneva, World Health Organization; 2016.

[23] Gompertz, B., 1825. On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies. Philosophical Transactions of the Royal Society. 115: 513–585. doi:10.1098/rstl.1825.0026.

[24] Gotoh, J. Y., Shinozaki, K., & Takeda, A., 2013. Robust portfolio techniques for mitigating the fragility of CVaR minimization and generalization to coherent risk measures. Quantitative Finance, 13(10), 1621-1635.

[25] Graunt, J., 1662. Natural and Political Observations made upon the Bills of Mortality, London.

[26] Haenszel, W., 1950. A standardized rate for mortality defined in units of lost years of life.

[27] Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org or www.humanmortality.de (data downloaded on 2017/Sep/17).

[28] Inskip, H., Beral, V., Fraser, P. and Haskey, J., 1983. Methods for age-adjustment of rates. Statist. Med., 2: 455-466. doi:10.1002/sim.4780020404.

[29] J Cuzick, H Stewart, L Rutqvist, J Houghton, R Edwards, C Redmond, R Peto, M Baum, B Fisher, and H Host, 1994. Cause-specific mortality in long-term survivors of breast cancer who participated in trials of radiotherapy. Journal of Clinical Oncology 1994 12:3, 447-453

[30] J. W. Rutter, 2000. Geometry of Curves. Chapman & Hall/CRC.

[31] Keyfitz, N., 1966. Sampling variance of standardized mortality rates.

[32] Krokhmal, P., Palmquist, J., & Uryasev, S., 2002. Portfolio optimization with conditional value-at-risk objective and constraints. Journal of risk, 4, 43-68.

[33] Ledoit, O., & Wolf, M., 2003. Honey, I shrunk the sample covariance matrix.

[34] Linder, F. E., & Grove, R. D., 1943. Vital statistics rates in the United States, 1900-1940. US Government Printing Office.

[35] Little, R. J., 1988. Robust estimation of the mean and covariance matrix from data with missing values. Applied Statistics, 23-38.

[36] Makeham, W. M., 1860. "On the Law of Mortality and the Construction of Annuity Tables". J. Inst. Actuaries and Assur. Mag. 8: 301–310.

[37] Moss, S. E., Klein, R., & Klein, B. E., 1991. Cause-specific mortality in a population-based study of diabetes. American journal of public health, 81(9), 1158-1162.

[38] Nawrocki, D. N., 1999. A brief history of downside risk measures. Journal of Investing, 8, 9-25.

[39] O'brien, R. M., 2007. A caution regarding rules of thumb for variance inflation factors. Quality & quantity, 41(5), 673-690.

[40] Paul Velleman, Roy Welsch, 1981. Efficient Computing of Regression Diagnostics. The American Statistician. American Statistical Association. 35 (4): 234–242.

[41] Reulen RC, Winter DL, Frobisher C, et al., 2010. Long-term Cause-Specific Mortality Among Survivors of Childhood Cancer. JAMA. 2010;304(2):172–179. doi:10.1001/jama.2010.923

[42] Shang, C., Huang, X., & You, F., 2017. Data-driven robust optimization based on kernel learning. Computers & Chemical Engineering, 106, 464-479.

[43] Stubbs, R. A., & Vance, P., 2005. Computing return estimation error matrices for robust optimization. Axioma Research Papers, 1, 1-9.

[44] United Nations, Department of Economic and Social Affairs, Population Division, 2015. World Population Prospects: The 2015 Revision, DVD Edition.

[45] Vapnik, V., 2013. The nature of statistical learning theory. Springer science & business media.

[46] Velleman, P.F. and Welsch, R.E., 1981. Efficient computing of regression diagnostics. The American Statistician, 35(4), pp.234-242.

[47] Yerushalmy, J., 1951. A mortality index for use in place of the age-adjusted death rate. American Journal of Public Health and the Nation's Health, 41, 907-922.

[48] Yule, G. U., 1934. On some points relating to vital statistics, more especially statistics of occupational mortality. Journal of the Royal Statistical Society, 97(1), 1-84.

[49] Zhu, J., Hoi, S. C., & Lyu, M. R. T., 2008. Robust regularized kernel regression. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 38(6), 1639-1644.

Table 17: Lookup table for reference populations and their Q90 and slope (year 2000)

Reference Population	Reference Q90	Slope
WHO2000	62.1	-0.0516
US2000	68.4	-0.0675
Canada	67.8	-0.0650
UK	71.1	-0.0915
France	72.1	-0.0849
Germany	72.1	-0.0921
Norway	71.1	-0.0882
Greece	72.3	-0.0758
Australia	68.3	-0.0571
Japan	71.0	-0.0979

Appendix

Fact. Sample covariance matrix can be decomposed into two parts with the first part consisting of the first n principal components and the second part consisting of the rest $N - n$ principal components.

Proof.

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{T}(\sum_{i=1}^{N} \boldsymbol{X}_i \cdot \boldsymbol{P}_i^\top)^\top \cdot (\sum_{i=1}^{N} \boldsymbol{X}_i \cdot \boldsymbol{P}_i^\top)$$

$$= \frac{1}{T}(\sum_{i=1}^{n} \boldsymbol{X}_i \cdot \boldsymbol{P}_i^\top + \sum_{i=n+1}^{N} \boldsymbol{X}_i \cdot \boldsymbol{P}_i^\top)^\top \cdot (\sum_{i=1}^{n} \boldsymbol{X}_i \cdot \boldsymbol{P}_i^\top + \sum_{i=n+1}^{N} \boldsymbol{X}_i \cdot \boldsymbol{P}_i^\top)$$

$$= \frac{1}{T}(\sum_{i=1}^{n} \boldsymbol{X}_i \cdot \boldsymbol{P}_i^\top)^\top \cdot (\sum_{i=1}^{n} \boldsymbol{X}_i \cdot \boldsymbol{P}_i^\top) + \frac{1}{T}(\sum_{i=n+1}^{N} \boldsymbol{X}_i \cdot \boldsymbol{P}_i^\top)^\top \cdot (\sum_{i=n+1}^{N} \boldsymbol{X}_i \cdot \boldsymbol{P}_i^\top)$$

$$= \hat{\boldsymbol{\Sigma}}_n + \hat{\boldsymbol{\Sigma}}_{N-n}$$

where $\boldsymbol{X}_n$ is the score vector with respect to the nth principal component $\boldsymbol{P}_n$.

□

Fact. Given a semi-definite symmetric matrix $\mathbf{\Sigma}$, we have

$$cond(\mathbf{\Sigma}) \geq cond(diag(\mathbf{\Sigma}))$$

where $diag(\mathbf{\Sigma})$ is the diagonalization of matrix $\mathbf{\Sigma}$, and the conditional number $cond(\mathbf{\Sigma})$ is defined as the ratio of the maximum eigenvalue to the minimum eigenvalue of matrix $\mathbf{\Sigma}$.

Proof. Because $\mathbf{\Sigma}$ is symmetric, it has the following Spectral Decomposition:

$$\mathbf{\Sigma} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$$

where $\boldsymbol{Q}\boldsymbol{Q}^\top = 1$, $\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_N \end{bmatrix}$, and $\lambda_1, ..., \lambda_N$ are N eigen values of $\mathbf{\Sigma}$.

Therefore,

$$\mathbf{\Sigma} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$$

$$= \begin{bmatrix} q_{11} & \cdots & q_{1N} \\ \vdots & \ddots & \vdots \\ q_{N1} & \cdots & q_{NN} \end{bmatrix} \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_N \end{bmatrix} \begin{bmatrix} q_{11} & \cdots & q_{N1} \\ \vdots & \ddots & \vdots \\ q_{1N} & \cdots & q_{NN} \end{bmatrix}$$

$$= \begin{bmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{N1} & \cdots & a_{NN} \end{bmatrix} \begin{bmatrix} q_{11} & \cdots & q_{N1} \\ \vdots & \ddots & \vdots \\ q_{1N} & \cdots & q_{NN} \end{bmatrix} \tag{8}$$

where $a_{ij} = q_{ij}\lambda_j$.
Let $diag(\mathbf{\Sigma})$ denote the diagonalization of matrix $\mathbf{\Sigma}$, we have

$$diag(\mathbf{\Sigma}) = \begin{bmatrix} d_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & d_N \end{bmatrix} \tag{9}$$

Comparing (9) with (8), we have

$$d_i = a_{i1}q_{i1} + ... + a_{iN}q_{iN}$$

$$= q_{i1}\lambda_1 q_{i1} + ... + q_{iN}\lambda_N q_{iN}$$

$$= \lambda_1 q_{i1}^2 + ... + \lambda_N q_{iN}^2$$

.

Notice that $\boldsymbol{Q}\boldsymbol{Q}^\top = 1$, therefore

$$q_{i1}^2 + ... + q_{iN}^2 = 1, \, \forall i \in \{1, ..., N\}$$

Thus, d_i is the weighted average of $\lambda_1, ..., \lambda_N$.
Therefore, we have

$$\max\{\lambda_1, ..., \lambda_N\} \geq \max\{d_1, ..., d_N\}$$

$$\min\{d_1, ..., d_N\} \geq \min\{\lambda_1, ..., \lambda_N\}$$

$$cond(\mathbf{\Sigma}) = \frac{\max\{\lambda_1, ..., \lambda_N\}}{\min\{\lambda_1, ..., \lambda_N\}} \geq \frac{\max\{d_1, ..., d_N\}}{\min\{d_1, ..., d_N\}} = cond(diag(\mathbf{\Sigma}))$$

□