How to Live Longer:

Insights from Predicting Life Expectancy in the Chinese Elderly Population (75 yr+)

12.15.2019

Jiou Choi, Kelsey Majam, Kevin Tiankun Wang
STAT 471-402
Prof. Linda Zhao

Table of Contents

Executive Summary

Life expectancy is a crucial measure of health at a societal and individual level. To inform interventions that may improve longevity and policies that rely on life expectancy estimates, we implemented multiple statistical models to **predict the number of remaining years of life for senior Chinese citizens**. Our final model was a **LASSO-optimized linear mode**l which predicted remaining life years with an **R2 accuracy of 0.244** and a **mean squared testing error (MSE) of 6.55**. Variables that we found significantly improve life expectancy were the province, age, gender, physical activity, pet raising, marital status, regularity of heart rate, weight, motor ability, and cognitive ability.

While there is clearly room for improvement in terms of prediction accuracy, a major limitation of this study was the **skewness of the response variable**, since most respondents died early within our data. Thus, normality of the residuals was not met, since the longevity of those that survived the longest was difficult to predict. One of our major recommendations would be to gather more sample data on those who survive the longest.

However, since some variables were still found to be significant, there are some immediate next steps in the form of suggested interventions. Based on our findings, **greater investment into senior centers** and other public investment to encourage physical activity could significantly improve life expectancy.

Background

Life expectancy can be defined as the number of years of life remaining at a given age.[1] Predicting life expectancy is useful for 3 main reasons:

1. As a summary statistic, life expectancy is used to **measure a society's health and wellbeing**. It may serve as a proxy for the quality of health services, healthy behaviors within the population, and general socioeconomic conditions.[2]
2. In practice, life expectancy estimates are crucial for **many commercial and federal plans**. For instance, life expectancy is used to determine life insurance policies and also used for government budgeting of social benefits programs.[3]
3. Finally, as the **bottom-line indicator for health**, prediction of low life expectancies can be met by societal or personal interventions to improve health and longevity.[4]

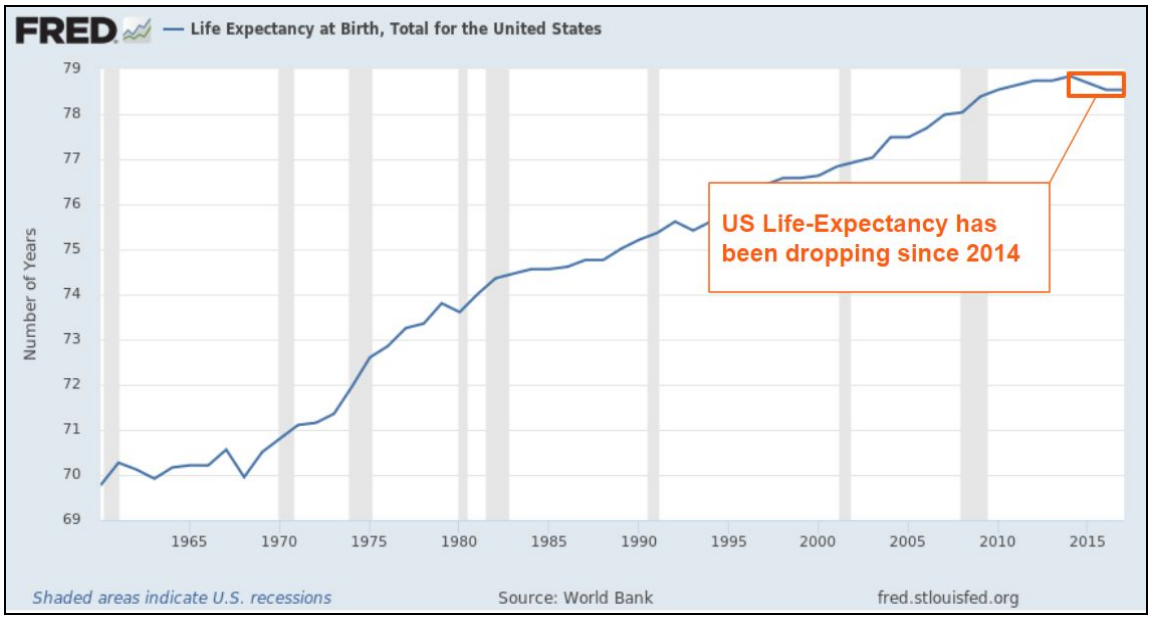
[1] Sullivan, Arthur O. and Steven M. Sheffrin. "*Economics: Principles in Action*." *Pearson Prentice Hall*. p. 473

[2] Ho, Jessica Y, and Arun S Hendi. "Recent Trends in Life Expectancy across High Income Countries: Retrospective Observational Study." *Bmj*, 2018.

[3] "Social Security." *Benefits Planner | Life Expectancy | Social Security Administration*, www.ssa.gov/planners/lifeexpectancy.html.

[4] Stiefel, Matthew C., et al. "A Healthy Bottom Line: Healthy Life Expectancy as an Outcome Measure for Health Improvement Efforts." *Milbank Quarterly*, vol. 88, no. 1, 2010, pp. 30–53.

A major concern for developing countries is that while life expectancy has increased dramatically in the last century, this decades-long linear trend has slowly stagnated. For instance, in the United States, at-birth life expectancy has recently been **decreasing by ~0-0.1 years annually since 2014**.[5,6]



Thus, in this study, we aim to create a statistical model that predicts **life expectancy**, defined as the **additional years of life remaining**. By identifying significant factors correlated with life expectancy, the insights gleaned from this study may be used to develop interventions that allow the elderly to live longer lives. Furthermore, the exploratory data analysis that we conducted may be compared to the general population in order to determine what factors allow people to reach an elderly age in the first place.

Summary of Data Used and Methods Used

The data used is from the **Chinese Longitudinal Healthy Longevity Study (CLHLS)** downloaded from the National Archive of Computerized Data on Aging.[7] CLHLS provides information on the health and quality of life of elderly aged 75 and older in 22 provinces of China from 1998 to 2014. The study follows the "oldest-of-the-old" over a series of 7 followup interviews. The respondents were asked about their health conditions, daily

[5] Uptin. "US Life Expectancy Has Been Declining. Here's Why." *CNBC*, CNBC, 9 July 2019, www.cnbc.com/2019/07/09/us-life-expectancy-has-been-declining-heres-why.html.
[6] "Life Expectancy at Birth, Total for the United States." *FRED*, 3 May 2019, fred.stlouisfed.org/series/SPDYNLE00INUSA.
[7] Zeng, et al. "Chinese Longitudinal Healthy Longevity Survey (CLHLS), 1998-2014." *Chinese Longitudinal Healthy Longevity Survey (CLHLS), 1998-2014*, Inter-University Consortium for Political and Social Research, www.icpsr.umich.edu/icpsrweb/NACDA/studies/36692.

functioning, and self-perception of their quality of life, and tested on their mental acuity and physical activities to determine health status. The respondents and their families also provided information regarding lifestyle, diet, medical needs, drinking/smoking habits, socioeconomic factors, and family life. Basic medical testing, such as blood pressure and heart rhythm measures were conducted. During the course of the study, most participants passed away and the age of death recorded.

Response Variable: Additional Years Lived

Based on our data, we predicted longevity by predicting the **additional years that the participants lived after the 1998 base study (add_years_lived)** using linear regression with backwards selection, LASSO-penalized linear regression, and random forest models. Age at death was not directly predicted due to potential illogical cases of predicting someone to die at an age younger than their current age (i.e. an 80 year old being predicted to die at 78). However, additional years of life can be added to the current age for a more meaningful prediction of age at death.

Key Findings and Conclusion

We found that several variables remained statistically significant across our models, including province, age, gender, physical activity, pet raising, marital status, regularity of heart rate, weight, motor ability, and cognitive ability. Many of these confirm conventional wisdom for a healthy life, including staying active both physically and mentally. However, the significance of being married in increasing life expectancy is confirms a rather new trend in longevity research.[8] Another rather novel finding is that having pets seems to significantly increase life expectancy, this is in line with a recently published research paper showing that dog ownership may improve cardiovascular health and overall survival.[9]

Limitations

There are many limitations to our study. First, the dataset only covers 22 provinces of China and the participants were largely of Han Chinese descent and all over the age of 75. Therefore, it may not be representative of non-Chinese populations nor applicable to those under 75.

Furthermore, we only accounted for the additional years lived for the very old, many of the participants began the study when they were ~100 years old. Therefore, there is significant survivorship bias. More importantly, the additional life years was not normally distributed, which may play a role in our relatively low prediction accuracy. We saw that in the normal

[8] Harvard Health Publishing. "Marriage and Men's Health." *Harvard Health*, www.health.harvard.edu/mens-health/marriage-and-mens-health.
[9] Kramer, Caroline K., et al. "Dog Ownership and Survival." *Circulation: Cardiovascular Quality and Outcomes*, vol. 12, no. 10, 2019

Q-Q plots, there were larger prediction errors for those who had higher life expectancies because they are not as well represented in the data.

Next, our model excluded health data of siblings or children, because of extremely heterogeneous data. However, with the growing attention to hereditary and genetic risk-factors for diseases, factoring in the health status of relatives is an important area for future study.

Finally, it is unfortunate that our data contained many missing values within categorical variables. While most of these factors were not deemed statistically significant, the interpretation of these correlations should not be extrapolated as we do not know why that data is missing.

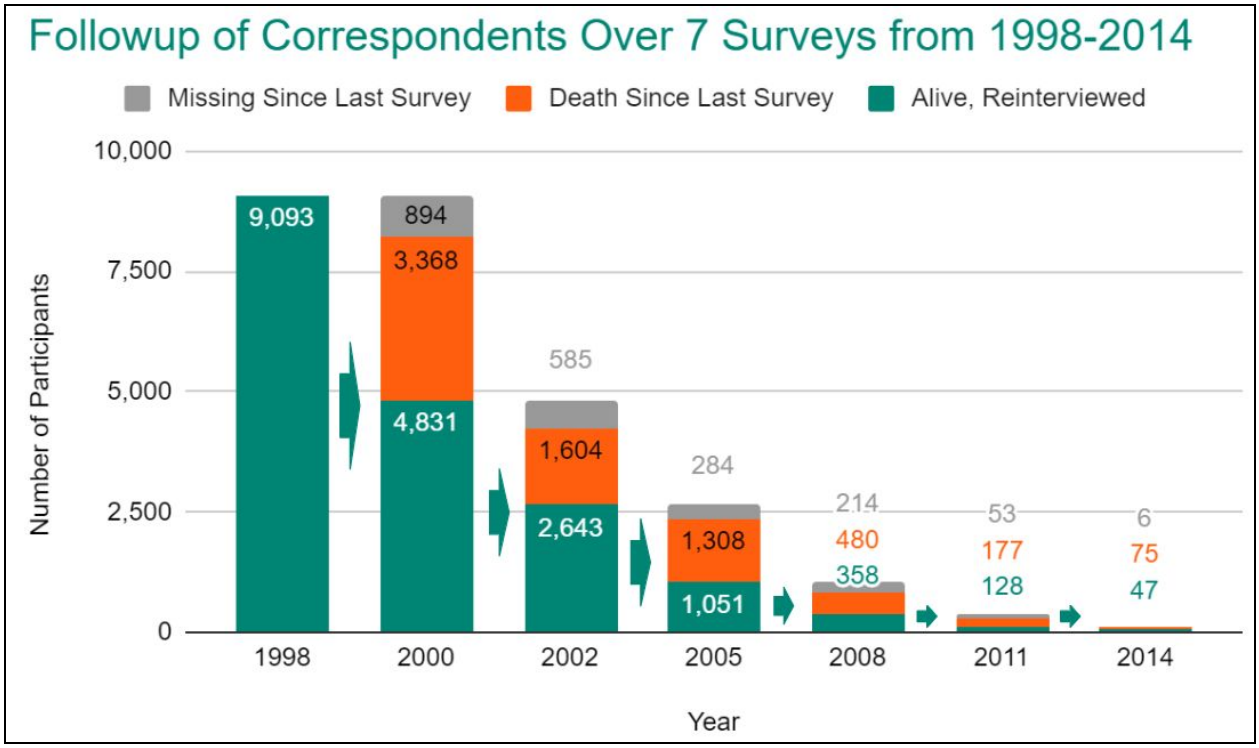
Our Analysis Process

To accurately predict the life expectancy of seniors and create a model that is interpretable, we began with a thorough reduction and cleaning of our extensive data set. The trimmed down dataset was then characterized through an exploratory data analysis. Finally, we used this dataset to implement 3 predictive models and evaluate each of them. The following sections will go into further detail about:

1. **Data Cleaning**
2. **Exploratory Data Analysis**
3. **Model 1: Backwards Linear Regression**
4. **Model 2: LASSO-penalized Linear Regression**
5. **Model 3: Random Forest Regression**
6. **Final Model Selection**

1. Data Cleaning

We began with the full dataset of **9093 observations** of **4740 variables** that was downloaded from the National Archive of Computerized Data on Aging.[10] However, because these 9093 observations were based on repeat surveys conducted in 1998, 2000, 2002, 2005, 2008, 2011, and 2014, there was significant loss of respondents due to death or inability to follow-up.



Since our variable of interest is the additional number of surviving years, we removed all participants who went missing and the participants who were still alive by 2014. This left us with **7010 observations**.

From looking through the variables across the years, we found that the 1998 interview questions and tests/measurements were the most comprehensive. Although some follow-up questions were asked in later years, there were a lot of missing values due to missing or deceased participants. Therefore, we decided to focus only on the most thorough set of metrics and answers from 1998 and disregard the follow-up interview material in later years, keeping only the death-related data. Removal of the follow-up questions left us with **863 variables**.

[10] Zeng, et al. "Chinese Longitudinal Healthy Longevity Survey (CLHLS), 1998-2014." *Chinese Longitudinal Healthy Longevity Survey (CLHLS), 1998-2014*, Inter-University Consortium for Political and Social Research, www.icpsr.umich.edu/icpsrweb/NACDA/studies/36692.

Next, we mutated the year of death to find the age at death by adding the difference between year that death was recorded and 1998 to the true age at 1998. We removed the observations with no confirmed age at death. We found the additional years lived by subtracting the age at 1998 from the age at death. We then removed all other variables from 2000-2014 except age at death and additional years lived, leaving us with **6993 observations** and **195 variables**.

After that, we began to mutate and condense predictor variables. First, we mutated and grouped 25 variables which were **tests of mental acuity**. The "mental_food" variable, which was the number of foods that the respondent was able to recall in a minute, was binned so that those who recalled 3 or less got a score of 0, those that named between 4 and 14 received a score of 0.5, and those than recalled more received a score of 1. All the other mental acuity predictors were mutated so that the respondent received a score of 1 if they answered correctly and 0 if they answered incorrect or they could not answer. Among the predictors, there were several tests that measured similar aspects of mental acuity. For example, whether a respondent can copy a drawing, fold a piece of paper, and pick up a paper on the floor all test for dexterity. We thought that the results of all similar tests together would be more informative than the independent results. Therefore, we grouped the variables into **"mental_recall"**, **"mental_verbal"**, **"mental_math"**, and **"mental_dexterity"** and assigned scores equal to the mean of all the test scores in that category. These groupings decreased our total variables to **173 variables**.

For data about **drinking, smoking, exercise, and labor habits**, we condensed the age that the habit began and the age that the habit ended into the duration of the habit until 1998. We also removed the data about whether the participant quit the habit since this is captured under whether the participant currently performs the habit. This transformation and the removal of the original variables reduced our variable count to **166 variables**.

We then mutated the **disease related variables**. We combined the disease and disease-related disability variables and relabeled all the levels to be more clear. For instance, instead of having a variable for whether the participant had hypertension and a second one for how disabling the hypertension was, we only kept the second variable and created a category of "**no disease**". This brought our penultimate dataset to **140 variables**.

For **all continuous variables** in the penultimate dataset, the original data entered in -1 for not applicable, and either 88, 888, 99, or 999 to indicate missing data. These were all changed to NA in order to not give numeric significance to their values.

Finally, for the last cleaning step, continuous variable columns and rows with greater than 10% of NAs were removed. The NAs for continuous columns were imputed with the median of each column.[11] Although doing this form of imputation could lead to less

[11] Laaksonen, Seppo. "Imputation Methods for Single Variables." *Survey Methodology and Missing Data*, 2018, pp. 171–195.

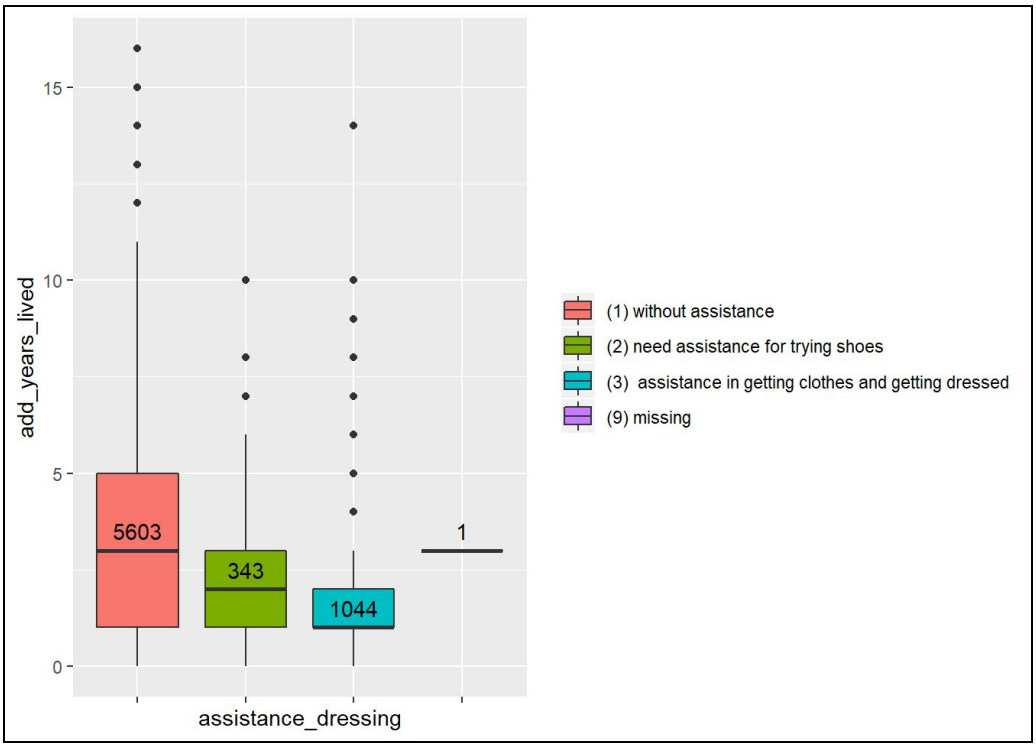
variable significance, this tradeoff was taken because there was a high number of columns such that listwise deletion would remove too many samples from our dataset. Finally, all categorical variables were converted into factors. Thus, through the data cleaning, our initial dataset of **9093 observations of 4740 variables** was narrowed down into **6991 observations of 127 variables**.

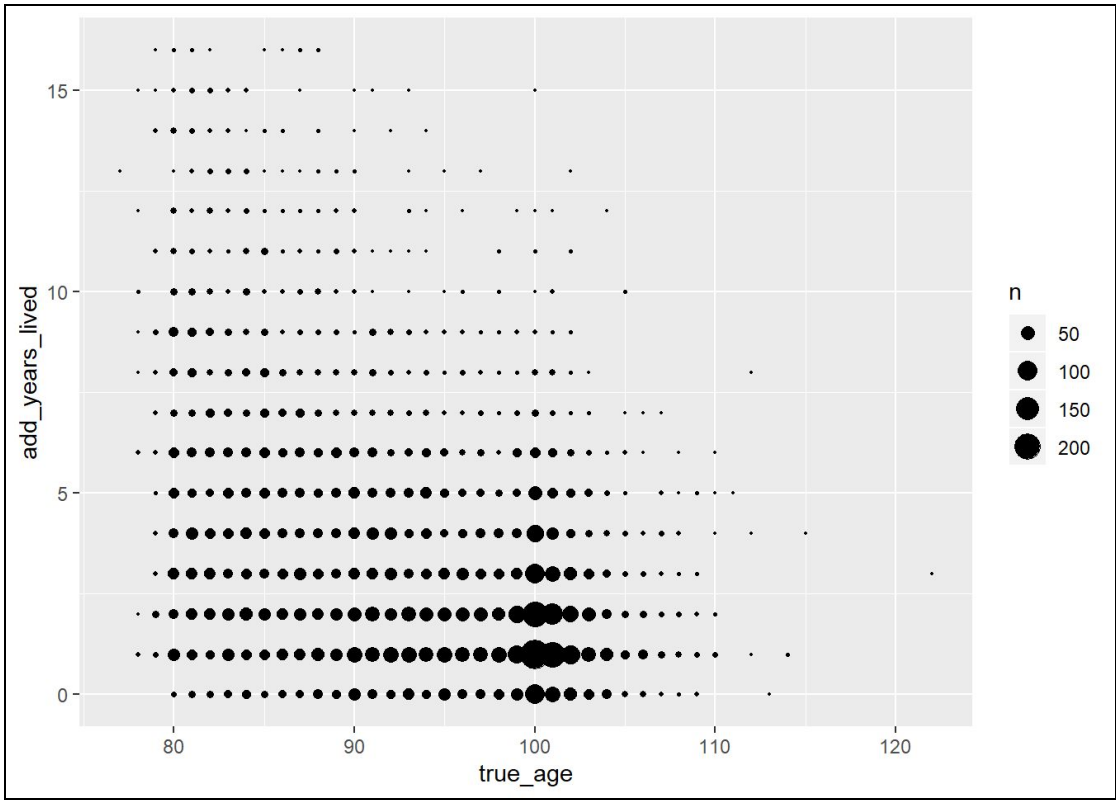
Unfortunately, for categorical variables, a large amount of **"(9) missing"** was present as a value within our data. Because of the large number, we did not wish to impute these missing data, because it could introduce significant additional bias into our model. As a result, we decided to leave it as a category. Generally, the **"(9) missing"** was not found to be a significant predictor, and the interpretations for these correlations should not be extrapolated.

2. Exploratory Data Analysis (EDA)

During the data cleanup, some initial exploratory data analysis was conducted to inform the cleanup process. However, this analysis was repeated and the final results can be found in the **EDA.html document**. For each categorical variable, the summary shows the frequencies of each category, and a boxplot shows the distribution of the categories with **add_years_lived**. For each continuous variable, the summary shows the values of each quartile and the mean. Scatterplots show the distribution of each continuous variable **add_years_lived**, and the size of the points correspond to the frequency of the distribution.

While how the explanatory variables were distributed will be talked at a high level within this section, the below plots are examples of the exploratory data analysis we conducted. **All 127 plots are available for review in the EDA.html document**.


add_years_lived
15
10
5
0
5603
343
1044
1
assistance_dressing
(1) without assistance
(2) need assistance for trying shoes
(3) assistance in getting clothes and getting dressed
(9) missing


add_years_lived
15
10
5
0
80
90
100
110
120
true_age
n
50
100
150
200

Demographics: We created boxplots for demographic variables to observe variations in additional years lived. There were no notable differences in additional years lived for province, province of birth, or gender. For ethnicity, Zhuang, Korean, and Mongolian people seem to live longer, but the number of observations in these categories are much smaller than that of the majority Han group, so the difference may not be significant. From the urban/rural birth boxplot, we can see that those born in an **urban setting** live for more years, potentially due to growing up in a more developed setting and having more resources.

The scatterplot of additional years lived vs. true age shows that most of the respondents were around 100 years old in 1998 base year. The size of the circles indicate that survival drops after about 6 additional years across ages. The general trend is that the number of additional years lived **decreases with age**, which is expected.

Psychological: Psychology and sense of wellbeing appeared to play a role in predicting additional years lived, as shown by the constructed boxplots for self-perceived lifestyle and health metrics. Those who felt **healthier, more optimistic, and more clean** had greater median number of additional years lived. Those who always or often **felt anxious or lonely** had a lower median additional years lived. Those with **greater independence and sense of purpose** and those who more frequently **felt as happy as when they were younger** lived for a greater median number of additional years.

Mental: The composite scores of **mental_recall**, **mental_verbal**, **mental_math**, and **mental_dexterity** seem to be linked to additional life years lived. Those that performed worse on mental tests generally had a lower life expectancy, as seen through the respective boxplots. Furthermore, the boxplot of reasons that the respondents were unable to answer the mental acuity test questions shows that those who were **paralyzed, those that had cognitive impairments, and those that were unable to participate due to illness** had a lower median number of additional years lived.

Diet: Participants who ate **wheat-based foods** as their **staple_food** appeared to have a longer life expectancy than those that ate rice or corn based meals. Other boxplots follow the conventional wisdom of a well-balanced diet: those that ate **fruit, vegetables, meat, and fish everyday** appear have a longer median life expectancy. Additionally, those that **drink tea** and **eat garlic frequently** also appear to have more additional life years to live.

Habits: Our habit data confirm some expected results: those who engage in **routine physical activity** such as housework, fieldwork and gardening almost everyday have higher average additional years lived. But it appeared that hobbies did not need to be purely physical to be beneficial: those who **read, engaged with pets, played mahjong, and listened to the radio** all had higher median life expectancy than those who did not. One result that was unexpected was that those who were still smoking in 1998 also had

higher median life expectancy; however, this may be due to bias since those who were in poor health were forced to quit smoking.

Independence: Participants who could **dress, use the restroom, use transportation, feed, and remain continent by themselves** had higher median additional years to live, as expected.

Socioeconomic: Our data shows that those who can **rely on themselves or their spouses** for finance generally have a higher life expectancy as shown by both the finances and caregiver boxplots. Indeed, being currently married and **living with the spouse** increases the median number of additional years lived. Surprisingly, having adequate medical services or being hungry as a child didn't have an effect on median additional years of life.

Family: Having a mother that's still alive added a large amount of median additional years of life; however, very few participants were in this category so the result may not be significant. Other family factors such as the number of siblings, birth order, and number of children did not seem to be noticeably different for additional years of life.

Health: **Visual function**, **having false teeth**, and being able to **use chopsticks** were linked with higher life expectancy, possibly because these participants are better able to function. **Systolic blood pressure, diastolic blood pressure, and heart rate** were interesting because it appears that those that live the longest have intermediate values, rather than too high or too low. **Functional tests** such as being able to place both hands behind the neck or picking up a book from the floor were also liked to greater life expectancy. In terms of diseases, participants who had the disease but were **not affected by disabilities** had greater median life expectancies than those who had the disease and it caused either mild or severe disability. However, because these degrees of freedom were small, we do not expect to see disease data to be extremely significant in our final model. Finally, having **hearing without aid and being able to undergo a physical check** were associated with increased median life expectancy.

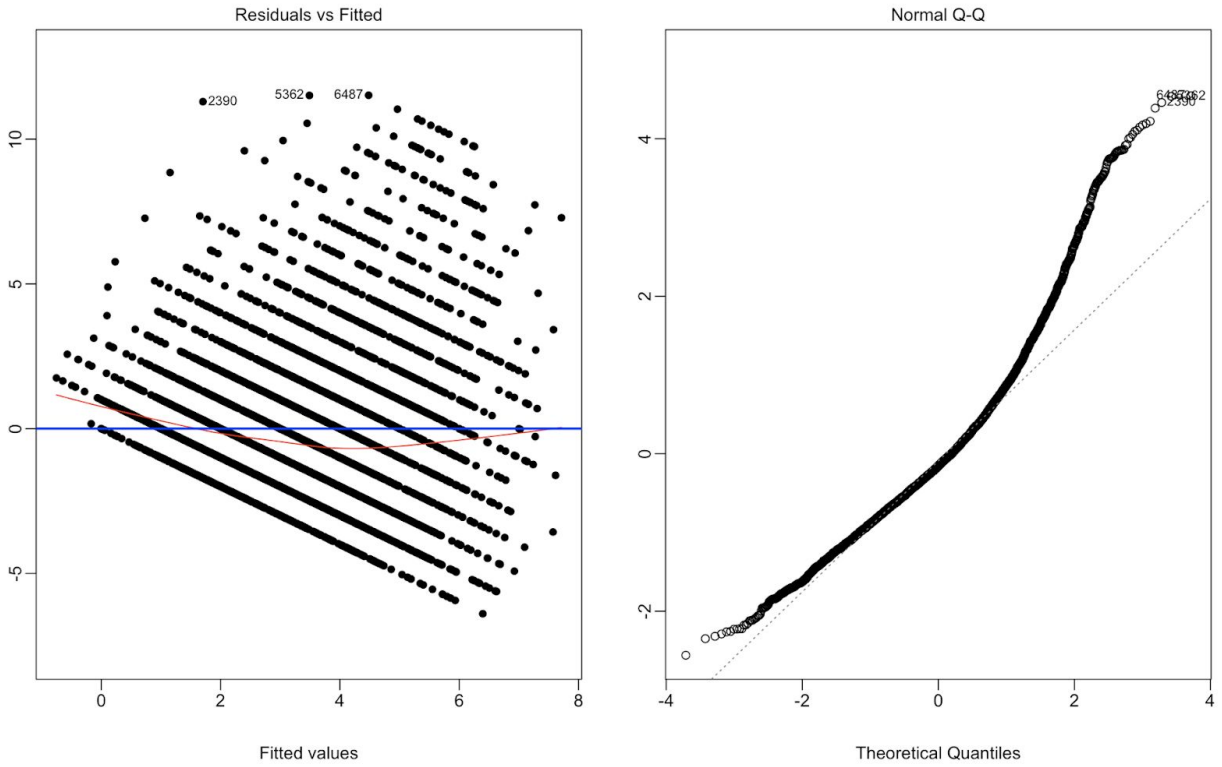
3. Model 1: Backwards Elimination Linear Regression

We started with an exhaustive multivariable linear regression with all 126 predictors (**fit0.train**). We ran an Anova test on the model to examine the significance of the variables. Due to the large number of variables, we removed at once all the variables with p-value greater than over equal to 0.9, then updated fit0.train to exclude those variables (**fit1.train**). We repeated these steps, removing the variables with the highest p-values and updating the model, until we had a final model with **13 variables**, all **significant at the 0.01 level**.

The final thirteen variables were **province, true age, gender, the frequency of vegetable consumption, whether they currently exercise, whether they currently do housework, whether they participate in raising pets, marital status, regularity of**

heart rhythm, weight, whether they can pick up a book from the floor, disability from heart disease, and the score from the math tests. From the values and the significance of the betas, females seem to have more years to live compared to males, and additional years lived decreases with the respondent's age at the 1998. Those that occasionally or almost never eat vegetables have a lower life expectancy than those that eat vegetables more frequently. Those that do not exercise, never do housework, never participate in pet care, widowed, and have an irregular heart rhythm have lower expected additional years to live. Life expectancy is positively correlated to the respondent's weight, which makes sense for a developing nation like China where obesity and weight-related chronic illnesses are not yet an issue. Those that can pick up a book from the floor while standing, do not have heart disease, and scored higher on the math tests have greater number of years to live.

The training error for the backward-elimination linear regression was **6.38 years**, while the testing error was **6.58 years**. As the maximum additional years lived was 16, a testing error of 6.58 years shows that this model does not perform very well. This is likely due to the poor prediction of long-surviving participants (see **normal Q-Q plot** below) because they were not well represented in the data. Still, the testing error was not much larger than the training error, meaning that the training data was representative of the testing data and that the model was not overfitted.



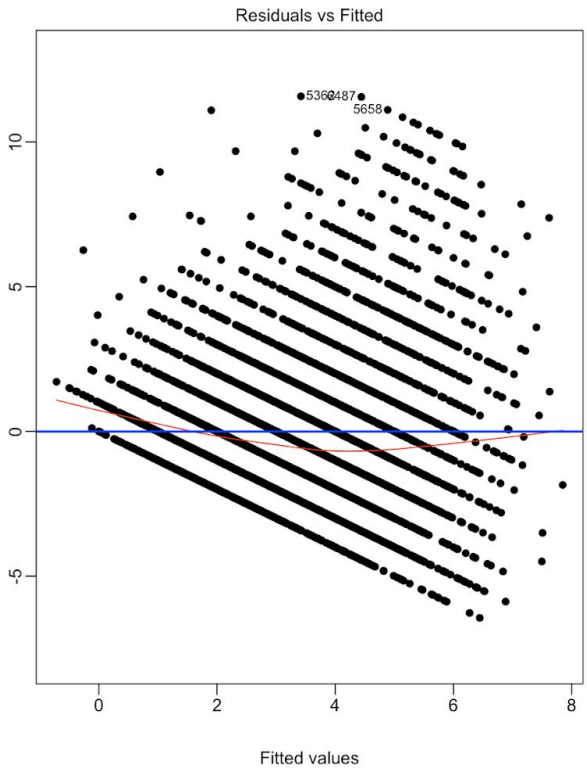
4. Model 2: Elastic Net Linear Regression

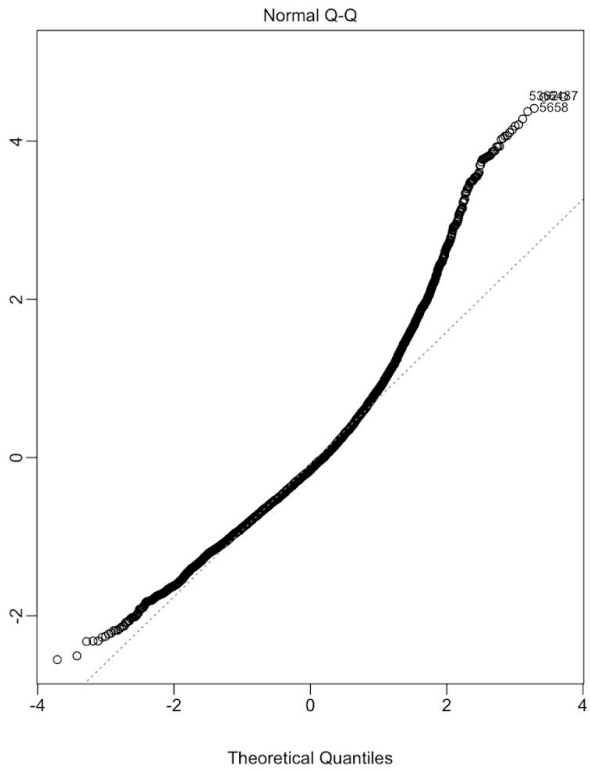
As we have many variables, we decided to run elastic net to narrow down the number of variables for a linear model. We constructed a matrix for all the predictors and ran an elastic net with nfold = 100 cross validation and alpha of 0.99. At lambda 1se, there were **18 variables with non-zero coefficients** (or variables with levels with non-zero coefficients). We refit these 18 variables into a regular linear regression, conducted an Anova test, and removed variables with p-value greater than 0.1. We then updated the regression with those variables removed, and repeated the process until all of the variables were significant at the 0.01 level.

The final model (fit.lasso) has **12 variables, most of which are the same as those in the backward elimination linear model**. The only differences are that the lasso model does not have the variables frequency of vegetable consumption and disability from heart disease, and instead has assistance with continence as a variable. The summary shows that those who do not need assistance with continence are expected to live for longer than those who are incontinent or have occasional accidents. The betas for the other variables are different in value but have the same signs as the previous model.

The residual plot shows that the linearity assumption is roughly met. However, the QQ-plot indicates that the normality assumption is not met at the upper extremes. Again, this may be because there are not many observations for the higher values of additional years lived.

The training error for this model is **6.4075 years**, and the testing error is **6.5462 years**, which are both lower than the errors for the backward elimination model.





5. Model 3: Random Forest Regression

We elected to run a random forest regression on our dataset in order to build several random deep trees at a time to bag in hopes of minimizing testing error. We tuned the parameter ntrees first by plotting ntrees by the out of bag error. An **ntrees of 250** was selected from the resulting graph as a substantial number of trees within the forest to help minimize testing error. We then selected for the parameter mtry through the loose formula of p/3 to give us an **mtry = 30**. With these chosen parameters, we ran our random forest model of 250 trees with 30 variables tried at each split.

The resulting random forest model (fit.rf) provides a training error of **7.1133 years**. While this is higher than the training error from the other two models, the random forest model succeeds in providing a slightly lower testing error of the three models (**6.4949 years**). Additionally, random forest presents us with the opportunity to utilize validation data in order to calculate validation error (**6.2177 years**).

6. Model Evaluation and Final Model Selection

Model	Type	R^2	MSE Testing	Significant Variables **(Bolded are unique to Model)**
1	Backwards Elimination Linear Regression	0.247	6.57	province, true age, gender, **the frequency of vegetable consumption**, whether they currently exercise, whether they currently do housework, whether they participate in raising pets, marital status, regularity of heart rhythm, weight, whether they can pick up a book from the floor, **disability from heart disease**, and the score from the math tests
2	Elastic Net Linear Regression	0.244	6.55	province, true age, gender, **whether they are incontinent**, whether they currently exercise, whether they currently do housework, whether they participate in raising pets, marital status, regularity of heart rhythm, weight, whether they can pick up a book from the floor, and the score from the math tests
3	Random Forest Regression	N/A	6.49	N/A

Conclusion / Final Recommendations

Our final three models are relatively comparable since the MSE only differs by very modest amounts.

Based on the above details of each model, we have chosen the final model to be the **Elastic Net Linear Regression model**. While it does not have a larger R^2 value than the Backwards Elimination Linear Regression model, it does provide a **lower testing error**. We acknowledge the lowest testing error from the Random Forest Regression model, yet due to the nature of the model, we cannot use it to identify the most significant variables as we can with the other two models.

Since the Elastic Net Linear Regression model is **easy to understand**, it is useful for suggesting actionable interventions. For instance, we see that exercise and housework are significantly positively associated with higher life expectancy. Thus, we can propose to **invest in senior centers or senior physical activity classes in order to improve their longevity**.

Appendix

Description of Variables

province: province of residence

true_age: age at the time of interview in 1998

gender: gender of the respondent

ethnicity: ethnicity of the respondent

province_birth: province the respondent was born in

urban.rural_birth: whether the respondent was born in an urban or rural area at the time of birth

self_QOL: self-perceived quality of life

self_health: self-perceived health

self_optimism: whether the respondent tends to look on the bright side of things

self_ clean: whether the respondent keeps their belongings and surroundings neat and clean

self_anxiety: how often the respondent feels anxious or fearful

self_lonely: how often the respondent feels lonely or isolated

self_independent: how often respondent feels that they can make their own decisions

self_useless: how often the respondent feels useless with age

self_happyAsYoung: how often the respondent feels as happy as they were when they were younger

staple_food: type of grain the respondent eats as a staple

amount_of_staple_food: amount (in liang) of the staple food the respondent eats per day

fruit_freq: how often the respondent eats fruit

veg_freq: how often the respondent eats vegetables

meat_freq_age_60: how often the respondent ate meat at age 60

meat_freq: how often the respondent eats meat

fish_freq_age_60: how often the respondent ate fish at age 60

fish_freq: how often the respondent eats fish

egg_freq_age_60: how often the respondent ate eggs at age 60

egg_freq: how often the respondent eats eggs

bean_freq_age_60: how often the respondent ate beans at age 60

bean_freq: how often the respondent eats beans

preserve_veg_freq_age_60: how often the respondent ate preserved vegetables at age 60

preserve_veg_freq: how often the respondent eats preserved vegetables

sugar_freq_60: how often the respondent ate sugar at age 60

sugar_freq_60.1: how often the respondent eats sugar

tea_freq_60: how often the respondent drank tea at age 60

tea_freq: how often the respondent drinks tea

garlic_freq_age_60: how often the respondent ate garlic at age 60

garlic_freq: how often the respondent eats garlic

water_source_child: the source of the water the respondent drank from as a child

water_source_to: the source of the water the respondent drank from at around age 60

water_source: the source of the water the respondent drinks from now

smoke: whether the respondent smokes now

freq_smoke: how often the respondent smoked/smokes

alcohol: whether the respondent drinks alcohol now

alcohol_type: what type of alcohol the respondent drinks

freq_alcohol.1: how often the respondent drank/drinks alcohol

exercise: whether the respondent exercises now

labor: whether the respondent does physical labor now

housework: how often the respondent does housework

fieldwork: how often the respondent does fieldwork/grows vegetables

garden: how often the respondent does gardening

read: how often the respondent reads

pets: how often the respondent actively participates in raising domestic animals/pets

cards_mahjong: how often the respondent plays mah-jong or cards

tv_radio: how often the respondent watches TV or listens to the radio

religion: how often the respondent participates in religious activities

assistance_bathing: whether/how much assistance the respondent needs to bathe

assistance_dressing: whether/how much assistance the respondent needs to dress

assistance_restroom: whether/how much assistance the respondent needs in using the restroom

assistance_transportation: whether/how much assistance the respondent needs in moving around

assistance_continence: how often the respondent has accidents

assistance_feeding: whether/how much assistance the respondent needs to eat

schooling_yr: years of schooling the respondent received

occupation: main occupation of the respondent before the age of 60

primary_finance: main source of financial support

other_finance: other sources of financial support

marital status: current marital status

num_marriage: number of times that the respondent was married

age_marriage1: age of respondent at first marriage

status_marriage1: status of the respondent's first marriage

quality_marriage1: quality of the first marriage

spouse_schooling_yr: years of schooling of the respondent's latest spouse

spouse_occupation: main occupation of the latest spouse before age 60

caregiver: the main caregiver when the respondent gets sick

adequate_medical_service: whether the respondent was able to get adequate medical service when he/she was sick recently

adequate_ms80: whether the respondent was able to get adequate medical service around age 80

adequate_ms60: whether the respondent was able to get adequate medical service around age 60

adequate_ms_child: whether the respondent was able to get adequate medical when he/she was a child

hungry_child: whether the respondent often went to bed hungry as a child

mother_alive: whether the respondent's mother is alive or not

father_alive: whether the respondent's father is alive or not

father_occupation: father's main occupation before the age of 60

num_sibs: number of siblings

birth_order: birth order of the respondent

num_child: number of children the respondent has

visual_function: whether the respondent can see and whether the respondent can distinguish a break in the circle

num_teeth: number of natural teeth the respondent has

false_teeth: whether the respondent has false teeth

use_chopsticks: whether the respondent is able to use chopsticks to eat

handedness: which hand the respondent usually uses for eating

systolic: systolic blood pressure

diastolic: diastolic blood pressure

heart_rhythm: whether the heart rhythm is regular or irregular

heart_rate: heart beats per minute

acromion_process_styloideus: the length of the radius bone of the arm

right_knee_to_floor: the height from the respondent's right knee to the floor

hand_behind_neck: whether the respondent can put right, left, or both hands behind their neck

hand_behind_back: whether the respondent can put right, left, or both hands behind their lower back

stand_from_chair: whether the respondent can stand up from sitting in a chair

weight: the respondent's body weight

book_from_floor: whether the respondent can pick up a book from the floor

num_ill_past2yr: number of times the respondent has suffered from a serious illness in the past two years

hypertension_dis: disability, if any, from having hypertension

diabetes_dis: disability, if any, from having diabetes

heart_disease_dis: disability, if any, from having heart disease

stroke_dis: disability, if any, from having a stroke

glaucoma_dis: disability, if any, from having glaucoma

cancer_dis: disability, if any, from having cancer

prostate_tumor_dis: disability, if any, from having prostate tumor

gastric_ulcer_dis: disability, if any, from having gastric ulcer

parkinson_dis: disability, if any, from having parkinson's disease

bedsore_dis: disability, if any, from having bedsores

other_chron: other chronic illnesses, if any

other_chron_dis: disability, if any, from having other chronic illnesses

hearing: whether the respondent was able to hear throughout the interview

physical_check: whether the respondent

reason_nophys: the reason why the respondent was not able to complete the physical check

add_years_lived: additional number of years lived beyond 1998

mental_recall: combination (average) of the scores of mental tests, including: 1) identifying time of day, 2) identifying the current month, 3) identifying the date of the mid-autumn festival, 4) identifying the season, 5) identifying the name of the county, 6) naming as many food as they can in one minute, 7) naming a pen, and 8) naming a watch

mental_ verbal: combination (average) of the scores of mental tests, including: 1) repeating the word "table", 2) repeating the word "apple", 3) repeating the word "clothes", 4) the number of attempts of repeat the words, 5) repeating the word "table" a second time, 6) repeating the word "apple" a second time, 7) repeating the word "clothes" a second time, and 8) repeating a sentence

mental_math: combination (average) of the scores of mental tests, including: 1) calculating 20 - 3, 2) calculating 20 - 3 - 3, 3) calculating 20 - 3 - 3 - 3, 4) calculating 20 - 3 - 3 - 3 - 3, and 5) calculating 20 - 3 - 3 - 3 - 3 - 3

mental_dexterity: combination (average) of the scores of mental tests, including: 1) taking paper using right hand, 2) folding a paper, and 3) putting paper on the floor

smoke_duration: the number of years the respondent smoked

alcohol_duration: the number of years the respondent drank alcohol

exercise_duration: the number of years the respondent exercised

labor_duration: the number of years the respondent did physical labor

Summary of Variables

See all in the EDA html document.

Boxplots and Scatterplots of Variables

See all in the EDA html document.