Human Lifespan is Well-Modeled by Linear Regression

Sergio E. Garcia Tapia

March 28, 2025 (last updated April 30, 2025)

1 Introduction

It has been suggested that the maximum human lifespan is fixed at around 122 years, and is subject to natural constraints [Bla21]. In contrast, the average life expectancy of a newborn increased from 32 years in 1900 to 71 years in 2021; on average, humans are living longer [Dat+23].

As life expectancy increases, it is natural to ask why we should care. One idea is that humans inherently strive for survival, so we naturally look for ways to live longer. Another answer is that a longer life brings more opportunities to nurturing social relationships with family and friends, or pursuing career and education goals. To what extent, then, can people influence their longevity to enable them to fulfill these life goals? The Danish Twin Study established that only 20% of how long the average person lives is dictated by our genes, whereas the other 80% is dictated by our lifestyle [BS16]. While on a quest to uncover aspects of lifestyle that increase longevity, Dan Buettner from National Geographic discovered regions of high longevity known as blue zones [BS16]. Blue zones are regions that consist of a large number of centenarians (people aged over 100) whose longevity is attributed to healthy lifestyles.

There are 9 primary lifestyle habits that people in blue zones adhere to which contribute to slow-aging. These factors can be summarized as: being in environments that naturally encourage movement; having a meaningful life purpose; having effective routines to cope with stress; eating until 80% full and eat less later times of the day; eating beans and little pork; drinking alcohol moderately; feeling a sense of belonging by being part of some community; prioritizing loved ones, such as family and friends; surrounding yourself with groups of people that encourage healthy behaviors [BS16]. A study by

Harvard's School of Public Health further strengthens the case that a healthy lifestyle can increase longevity. The study considered the influence of 5 factors similar to the ones we considered for blue zones, such as maintaining healthy eating patterns, not smoking, getting 3.5 hours of vigorous physical activity, drinking alcohol in moderation, and maintaining a "normal" weight. The study estimated that women at age 50 who did not adopt any of these 5 healthy habits were estimated to live until 79, whereas those who did were estimated to live to 93.1 years [Hic18].

A life insurance company may want to predict lifespan to maximize their profit from a customer's life insurance; on the flip side, a customer may use it to decide whether getting life insurance is worth it. A home buyer may weight the cost of buying a home against how many years they expect they will be able to live in it. Economists may want to understand how general trends in lifespan may impact the economy. For example, a home buyer of age 40 considering committing to a 30 year home payment may want to know if they will live long enough to enjoy their home for at least another 20 years past that. If a model predicts they will age to 95, it ought to be accurate to within about 5 years for it to be useful. Perhaps after learning this prediction, the home buyer might adjust their lifestyle to help them attain their goal. In this paper, I explore the extent to which specific factors can lengthen a person's expected lifespan.

2 Understand Reality via Models

In this paper, the phenomenon of interest is human lifespan. A phenomenon is a naturally occurring event whose manifestation can be observed. Phenomena range from natural disasters such as earthquakes and hurricanes to population growth and change in stock prices and even to gravity and magnetism. Phenomena can be quite complicated, and their study often requires making general assumptions to simplify their analysis. The result of this simplification is called a model.

A model is an abstraction, or an approximation, or a simplification, of reality. The goal of a model is to remove elements that affect a phenomenon whose consideration may be infeasible or whose contribution may be relatively insignificant. The adage "an apple a day keeps the doctor away" may be thought of as a model for living a healthy life. Another one is "Sitting is the new smoking", which warns that a sedentary lifestyle is a risk factor for cardiovascular morbidity [BSC16]. A person who intends to lead a healthy life may heed these proverbs by exercising more and eating healthier foods, but may quickly realize a problem. How many apples should they eat? How much sitting is too much sitting? The problem is that models can be ambiguous; though the qualitative description is useful for everyday talk, our inability to quantify the effect of each component can make it difficult to apply them. If we are to use models to guide the choices we make, we need something more rigorous.

A mathematical model is a specialization of a model that aims to establish a precise relationship between two or more quantities. To create a mathematical model, it is necessary to identify responses, also known as outputs, and inputs, also known as factors, as well as ways to measure them. The Navier-Stokes equations, for example, are a mathematical model believed to aid in explaining and predicting breeze and turbulence. [Ins25]. Even without the availability of exact solutions, the model has been useful in navigation in boats in water and modern jet flight across the sky. It is interesting to note that in spite of the ingenuous principles that Navier-Stokes equations embody, they are still far from perfect. In his book, Silver shared that a mathematician once said “The best model for a cat is a cat” [Sil12]. The mathematician meant that no model is perfect, for missing any detail inevitably leads to some inaccuracy. It is also important to remember that a model, mathematical or not, is not reality. Thus, even though models are useful, we ought to be careful not to let them be the final word on the phenomenon that we are attempting to describe with them.

3 A Model for the Lifespan Phenomenon

Let us consider how we would make a mathematical model for predicting the lifespan of an individual. A phenomenon has a response or output, mathematically denoted as y, whose values belong to a response space, a set denoted $\mathcal{Y}$. In human lifespan, y is the number of years that a person lives from birth to death, which can be 67.4 years or 100.12 years, for example. This quantity can be measured accurately from a person’s birth and death certificates. Thus, the response space $\mathcal{Y} = \mathbb{R}$, the set of all real numbers. To devise an exact mathematical relationship involving lifespan y, we need to identify its causal drivers, $z_1, z_2, \ldots, z_t$, and an exact functional relationship t so that we may write

$$y = t(z_1, z_2, \ldots, z_t).$$

The causal drivers are the true causal input information of the phenomenon's response. Here are some possible causal drivers influencing the lifespan of an individual:

z_1 = Number of fatal accidents experienced

z_2 = Presence or absence of genes to combat deadly disease

z_3 = Ability to gather healthy foods

z_4 = Average daily rate of sustained mental stress

z_5 = Level of physical aptitude

z_6 = Grit and desire for survival and longevity

One can argue that these inputs may not include all that goes into living a long life, or that some of these are not in fact causal drivers. In general, we do not have the omniscience to know exactly which inputs directly cause the response of a phenomenon to change. However, even if we did, we face the challenge of how to precisely measure these causal drivers. For example, it is impossible to know how many accidents a person will experience throughout their life. Similarly, the mental stress that a person experiences is linked to their life experiences and circumstances, which are always changing in unpredictable ways. Even if we did know all the causal drivers and were able to measure them, we would be hard-pressed to establish an exact functional relationship t relating them to the response y, because there are uncountably-many ways to relate these quantities.

4 Approximating Causal Drivers with Features

To deal with the infeasibility of obtaining true causal information, we introduce quantities known as features. A list of features $x_1, x_2, \ldots, x_p$, also known as predictors or covariates, act as proxies to the unattainable true drivers which are the $z_1, z_2, \ldots, z_t$. By proxies we mean that these features stand for them when describing the phenomenon. The list of features $x_1, x_2, \ldots, x_p$ is said to belong to a covariate space $\mathcal{X}$, which consists of all the possible lists involving different values of these features. We can express y through a functional relationship in terms of the features, but we must account for the fact that the features do not carry the same information density as the true causal drivers. That is, whereas we have the exact equation $y = t(z_1, z_2, \ldots, z_t)$, we only have an approximation $y \approx f(x_1, x_2, \ldots, x_p)$ for some function $f : \mathcal{X} \to \mathcal{Y}$, or

$$y = t(z_1, z_2, \ldots, z_t) = f(x_1, x_2, \ldots, x_p) + \delta$$

where δ is a positive real number. The quantity δ is known as ignorance error, encapsulating the error incurred by expressing y in terms the features $x_1, x_2, \ldots, x_p$ that are not truly causal. The following is list of features that may affect lifespan:

$x_1 =$ Daily average number of steps walked each day, an integer

$x_2 =$ Daily average number of hours of sleep

$x_3 =$ Daily average number of 12-oz cups equivalent to a wine glass

$x_4 =$ Daily average number of cigars a person smokes

$x_5 =$ Weekly average number of hours of pleasurable social interactions

$x_6 =$ Current age of individual in years

We learned earlier how centenarians in blue zones heed the habits embodied by some of these features (and more), so it is reasonable to believe that they affect lifespan. The advent of smart watches makes measuring x_1 and x_2 realistic, which can be a good measure of a person's level of physical activity and aptitude and a way to help cope with mental stress, respectively. Admittedly x_3 may not be as accurate as possible because not all wines or alcoholic beverages have the same alcohol content. However, a person can take note of when they purchase a box or a single unit of bottles of cans of some drink, and maintain a diary with a headcount of when and how much they consumed. Similarly, cigars are sold in boxes of multiple units, so a person can track their purchases to know x_4, how many cigars they have smoked. For x_5, a person can refer to their extracurricular activities, such as outings for restaurants or dancing, video or board game sessions, or even time spent at home talking with friends and family; a calendar or a diary can help track this. Finally, x_6 may be important to account for the fact that as we age older, we may be less likely to continue to live, or may be more susceptible to certain health problems.

A reasonable age to start collecting information about a person's features is 21. In most countries, since in most countries a person of this age is old enough to buy alcohol or cigars. People at this age also tend to own smart phones or smart watches, which tend to automatically track walk and sleep activity. At age 35, we can start making predictions about how long said person will live. As the person ages, we continue tracking their features to make refined and more accurate predictions as time passes.

5 Stationarity

A reasonable question is whether a model we create will continue to be valid. change. A phenomenon is stationary if the same drivers or features that we use in a model today (and features) still influence the response in the same way when we observe the model (say, 10 years later) under the same functional relationship t or f. When the relationship changes, the phenomenon is said to not be stationary. A model for a non-stationary model is said to drift, because the model encapsulates assumptions whose validity may not hold as the phenomenon changes, rendering the model invalid.

In general, most phenomena are not stationary, and we might only be able to achieve stationarity if we have a closed system. In particular, human lifespan is not a stationary phenomenon. For example, though lifespan is thought to have a maximum value, humans might make a scientific discovery to extend it. Thus, if our model is built around the assumption that humans can live at most 122 years, it will need to be revised.

6 Supervised Learning

So far, we can express the human lifespan response as

$$y = f(x_1, x_2, x_3, x_4, x_5, x_6) + \delta$$

where we rely on $p = 6$ features. However, the truth is that we cannot hope to attain an analytical solution. Earlier we saw an example of this when we mentioned the Navier-Stokes equations, for which no exact solutions exist in spite of the ambitious effort of the mature theory of differential equations [Ins25]. Instead of giving up, we can adjust our expectations and settle for an empirical solution. An empirical solution is an approximate solution that can be obtained from historical data, and the framework for obtaining a solution in this way is called supervised learning. Another name for supervised learning is machine learning.

The framework of supervised learning consists of a set of training (also known as historical) data $\mathbb{D}$, which consists of known inputs and their corresponding correct outputs. It is supervised in the sense that an entity who has collected the data has gone through the trouble of ensuring the data is correct for the phenomenon. We then decide on set of candidate functions $\mathcal{H}$ that we believe well-approximate (or may accurately describe) the relationship between the inputs and outputs. Finally we combine these using an algorithm $\mathcal{A}$ that uses these to build a function that we can use to make predictions. In the following sections, we will discuss each of these components in more detail.

6.1 The Historical Data

Historical data refers to a set $\mathbb{D}$ of n data points that are measured by observing the phenomenon. We can organize the data points as ordered pairs and enumerate them

$$\mathbb{D} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$$

where y_k is the output for the kth observation, expressed as $y_k = f(\boldsymbol{x}_k)$. For us, $\boldsymbol{x}_k = \begin{bmatrix} x_{k,1} & x_{k,2} & x_{k,3} & x_{k,4} & x_{k,5} & x_{k,6} \end{bmatrix}^\top$, which identifies a person whose health habits are known, and who is known to have lived for y_k years from birth to death. In the framework of supervised learning, a large number n of data points is necessary to ensure the model can have a fair try at capturing different patterns and thereby accurately make predictions. If our intent is to predict human lifespan, then our n must be reasonably large given that the world has about 8 billion people [Cen24]. An important consideration is how easily such data can be obtained.

In measuring the human lifespan phenomenon, collecting each data point literally takes a lifetime; an observation is not complete until someone has died. We could gather data on the features of live people through a survey and collect their features as a vector $\boldsymbol{x} \in \mathcal{X}$, then when they have passed in the future we associate that vector with a lifespan $y \in \mathcal{Y}$. If we wanted to have data now, it might be expensive or impossible because of data protection laws. One crude way to gather some data about lifespans y would be to use Wikipedia, which often lists the birth and death dates of individuals (with accuracy subject to verification), but we may not find the information on the specific features $\boldsymbol{x}$ that we are interested for that individual. After all, not everyone tracks how many alcoholic drinks they intake every day, or exactly how many hours of sleep they get at night. Moreover, technology such as smart watches to track a person's daily footstep count was either not invented or widely unavailable years ago. Another consideration is that we may need to prefer more recent deaths to account for the fact that human lifespan is a non-stationary phenomenon.

In summary, although we want a large number n of data points, data collection can be challenging. If we started from zero data today, we would likely need about 150 years of data collection to construct a training sample. Thereafter, we would be able to use that data to train a model and to make predictions.

6.2 Hypothesis Set

The next part of the supervised learning framework deals with the infeasibility of an analytical solution. Simply put, f is and will always be unknown, and it is likely much more complex than we can handle. Since we seek an approximate solution, our next task is to conjecture a functional form that we believe may well approximate f. The symbol $\mathcal{H}$ denotes a set of candidate functions, where $h \in \mathcal{H}$ means that $h : \mathcal{X} \to \mathcal{Y}$, meaning h is a function with the same domain and range as f. Our hope is to find the best function $h^* \in \mathcal{H}$ that approximates f. Such a function would be related to y, t, and f by

$$y = t(z_1, z_2, \ldots, z_t) = f(x_1, x_2, \ldots, x_p) + \delta = h^*(x_1, x_2, \ldots, x_p) + \epsilon$$

Here, ϵ is known as noise, which accrues the ignorance error mentioned earlier, and a new error quantity known as misspecificaton error, expressed as $\epsilon - \delta = t - f$. To understand how misspecification comes about, let us consider an example. Suppose we define the following set of candidate functions:

$$\mathcal{H}_1 := \{\, w_0 + w_1 x_1 + w_2 + w_3 x_3 + w_0 x_4 + w_5 x_5 + w_6 x_6 \mid w_0, w_1, w_2, w_3, w_4, w_5, w_6 \in \mathbb{R} \,\}$$

Set $\mathcal{H}_1$ is the set of hyperplanes, which is a generalization of the idea of lines to higher dimensions. Such a choice is attractive because of the ease of interpretation of the results, and is appropriate if the data appears to have a "linear" pattern. On the other hand, suppose our colleague proposes the following candidate set:

$$\mathcal{H}_2 = \{\, w_0 + w_1 x_1 x_2 x_5 + w_2 x_3 x_4 + w_3 x_6^2 \mid w_0, w_1, w_2, w_3 \in \mathbb{R} \,\}$$

Our colleague reasons that people who smoke are also more likely to drink, so the multiplicative term $x_3 x_4$, which is known as a first-order interaction of features, accounts for this. Similarly, they argue that walking, sleeping, and interacting with friends compound on each other some way, leading to other lifespan gains. Finally, they mention that current age weights heavily in the prediction, so it should be squared, yielding a transformed feature. Putting aside whether our colleague's claims are true, this example demonstrates that there are many ways to design a candidate set. Not all candidate sets may be equally effective in approximating f. For example if the data has a parabolic pattern, fitting a line may not be a good idea. The misspecification error $\epsilon - \delta$ quantifies this "mismatch" between the candidate set and the data set $\mathbb{D}$ that presumably came from f. I think our colleague may be right in considering a hypothesis set with interactions. For example, studies shows that smokers tend to sleep less than non-smokers [Jae+12]. Thus, interactions between the features may improve our model. However, they can also make our model more complex to reason about. This trade-off between performance and interpretability is a recurring theme that plays an important role in the modeling enterprise. Going forward we will use $\mathcal{H}_1$, hereby referring to it as just $\mathcal{H}$, as our set of candidate functions, because it is easier to interpret.

6.3 Algorithm

Once we have measured data $\mathbb{D}$ in the real world and chosen a candidate set of functions $\mathcal{H}$, we need a way to obtain a function h from $\mathcal{H}$. The set $\mathcal{H}$ is infinite and we need a way to search the space efficiently to obtain an approximation to f. To do this, we employ an algorithm, denoted $\mathcal{A}$, which is a finite sequence of steps devised to solve a problem. We can view an algorithm as a function $\mathcal{A}(\mathbb{D}, \mathcal{H})$ that accepts a data set and a set of candidate functions and outputs an approximation $h \in \mathcal{H}$ to f. Earlier we said that h^* is the best approximation to f in $\mathcal{H}$, so naively we might hope for an algorithm to output h^*. Once again, though, we must temper our expectations; as we have emphasized, $\mathcal{H}$ is huge, uncountably-infinite as a matter of fact. Another way to think about it is that many functions do not attain all possible outputs in their co-domain, and the algorithm $\mathcal{A}$ that we use is no different. We will say that the best possible function that our algorithm can produce is $g \in \mathcal{H}$, related to y, t, f, and h^* by

$$y = t(z_1, z_2, \ldots, z_t) = f(x_1, x_2, \ldots, x_p) + \delta = h^*(x_1, x_2, \ldots, x_p) + \epsilon = g(x_1, x_2, \ldots, x_p) + e$$

Introducing g also introduces a a new type of error called estimation, given by $h^* - g = e - \epsilon$. Thus the quantity e, called the residual, contains accrues ignorance, misspecification, and estimation. absorbed by the residual error e that also happens to contain ignorance and misspecification. In a future section, we will discuss errors in more detail.

Many ingenuous algorithms have been developed by very smart people, each subject to certain constraints and weighing different parts of the supervised learning problem differently. For example, if the response space in our phenomenon was binary, such as it would be if we were modeling the presence or absence of a disease, then we might encode that as response in the set $\mathcal{Y} = \{0, 1\}$. One algorithm, known as perceptron, is applicable if the data set $\mathbb{D}$ satisfies a condition known as linear separability, which most data sets do not satisfy. Another algorithm, known as logistic regression, produces a candidate function g whose output is not in $\{0, 1\}$, but rather it outputs a probability that a person has the disease. In the case of human lifespan with a numeric response space $\mathcal{Y} = \mathbb{R}$, some plausible choices of algorithms are K-nearest neighbors (KNN) and ordinary least squares regression (OLS). In this paper, I will choose OLS with the linear candidate set $\mathcal{H}$ that we introduced earlier. The appeal of this choice is that we can compute an exact solution g using techniques from calculus and linear algebra. This is in contrast to other approaches for which no exact solutions exist, and instead an optimization algorithm is necessary.

7 Predictions

Before considering OLS, let us re-center ourselves by remembering why we are using supervised learning in the first place. We want to understand our phenomenon, or at least, we want to make predictions about the response of the phenomenon given a unit and their features. Suppose that Luis is 24-year old individual who works part-time with an interest in planning for retirement and in pursuing a college education. He determines that knowing how long he will be alive for will help in his planning. We offer Luis help, but we tell him that we need some information about him. After some initial hesitation to share personal information, we gather the following vector:

$$\boldsymbol{x}_{\text{Luis}} = \begin{bmatrix} 12042 & 7.3 & 0.1 & 0 & 15.7 & 24 \end{bmatrix}^\top$$

That is, Luis walks on average 12,042 steps each day, sleep on average 7.3 hours each night, drinks 0.1 glasses of wine each day on average, does not smoke at all, spends about 15.7 hours a week on average socialization with friends and family, and has lived for 24 years. Having applied our algorithm and obtained a function $g = \mathcal{A}(\mathbb{D}, \mathcal{H})$, we can use g to make a prediction. For example, suppose that our algorithm produced the vector

$$\boldsymbol{b} = \begin{bmatrix} 10 & \frac{1}{1000} & \frac{20}{8} & 5 & -10 & 3 & 24 \end{bmatrix}$$

Then our function g is given by

$$\hat{y} = g(\boldsymbol{x}) = 10 + \frac{1}{1000}x_1 + \frac{20}{8}x_2 5x_3 - 10x_4 + 3x_5 - 0.1x_6$$

The quantity $\hat{y}$ is the predicted number of years for a person whose feature as given by the feature vector $\boldsymbol{x}$. For Luis, we can compute this number explicitly using our made-up model g:

$$g(\boldsymbol{x}_{\text{Luis}}) = 85.492$$

That is, our model predicts that Luis will live for about 85 years. Luis seems a bit disheartened, sharing that he wanted to live longer, but he is also curious as to whether our prediction is accurate. We tell Luis that our model is an approximation, so he should not let it have the final word. We humbly also tell him that there may be better models out there. How can we ensure that our model is as good as can be? We need an error metric.

8 Types of Error

Before discussing error metrics, we need to understand discuss how errors are introduced in our modeling framework. The first type of error, ignorance error, arises from using features $x_1, x_2, \ldots, x_p$ as proxies for the the true drivers (casual inputs) $z_1, z_2, \ldots, z_t$ and a function f domain contains the possible values of these features instead of the true phenomenon function t:

$$y = t(z_1, z_2, \ldots, z_t) = f(x_1, x_2, \ldots, x_p) + \delta$$

The error due to ignorance accounts for the fact that the features do not contain the same information density as the true drivers. One way to deal ignorance is to introduce new relevant features to describe the phenomenon, thus increasing the number of features p per observation. The next approximation we made was to introduce h^* from some candidate set $\mathcal{H}$:

$$f(x_1, x_2, \ldots, x_p) + \delta = h^*(x_1, x_2, \ldots, x_p) + \epsilon$$

The quantity ϵ is known as the noise, and the difference $\epsilon - \delta = f - h^*$ is the misspecification error. We touched on this earlier, and it is a result of choosing a candidate set $\mathcal{H}$ that may not correctly capture the functional form of f. For example, if the points in $\mathbb{D}$, presumably produced by f (subject to some error due to δ) seem to follow a parabolic pattern, then a set of hyperplanes may accrue some misspecification error. One way to reduce misspecification is to choose a larger set of candidate functions. The last type of error is known as the residual error, and shows up when we obtain a modeling function g from our algorithm:

$$h^*(x_1, x_2, \ldots, x_p) + \epsilon = g(x_1, x_2, \ldots, x_p) + \epsilon$$

The residual contains the ignorance and misspecification error, and it adds a new type of error known as estimation error, given by $e - \epsilon = h^* - g$. The estimation error results from not being able to obtain the best possible function h^* out of $\mathcal{H}$. One way to reduce estimation error is to employ a better algorithm $\mathcal{A}$. Another way is to ensure that the features that we use are indeed relevant. For example, suppose that a colleague recommends that we use a new feature x_7 that represents the number of digits of π that a person knows. It is hard to believe that this affects lifespan at all. Suppose we are omniscient and know that it indeed has nothing to do with lifespan. If our model accounts for it and produces a non-negative weight associated with x_7, then we accrue estimation error. Estimation error can also increase if we decide to transform our features, perhaps by introducing a new feature that is the square of an existing feature. For example, if we had a single feature x_1 and believed the data set had a quadratic pattern relative to this feature, then we would create a transformed feature x_1^2 to allow for parabolic solutions out of our candidate set. This may lead to a decrease in misspecification error, but possibly an increase in estimation error.

Another way to reduce misspecification is to ensure we use the correct model. Earlier we favored $\mathcal{H}_1$ over $\mathcal{H}_2$ on the premise that $\mathcal{H}_1$ is more interpretable. However, if $\mathcal{H}_1$ does not perform well in comparison to $\mathcal{H}_2$, then we may want to choose $\mathcal{H}_2$ anyway. Model selection involves computing different predictions, possibly with the same or different algorithms or candidate sets, and measuring their predictive performance to see which one is more accurate. For example we would compute $g_1 = \mathcal{A}(\mathbb{D}, \mathcal{H}_1)$ and $g_2 = \mathcal{A}(\mathbb{D}, \mathcal{H}_2)$, and select the one whose error metric is smallest.

In the current setting of human lifespan, I expect ignorance error and misspecification error to be most prevalent because 6 features is likely too few, and a linear relationship between them is unlikely.

9 Error Metrics

Suppose we have computed a prediction function with our model

$$g = \mathcal{A}(\mathbb{D}, \mathcal{H})$$

How good is g at predicting? For example, if g is any good, then we might expect that for a unit $(\boldsymbol{x}, y)$ for which the response y is known, the value $g(\boldsymbol{x})$ might be close to $\hat{y} = g(\boldsymbol{x})$. In fact, the residual which we defined earlier is a good start point

$$e = y - g(\boldsymbol{x})$$

We can think of the residual as how far our prediction is from the actual response. The error is zero when the prediction is exactly equal to the response, and otherwise it may be positive or negative depending on whether it overshoots the response or not.

Given n data points $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$, we can compute n corresponding residuals $e_1, \ldots, e_n$. One error metric that we can define with these residuals is called the sum-of-squared errors:

$$SSE := \sum_{k=1}^{n} e_k^2$$

The SSE is useful, but it suffers from at least two problems: it inevitably grows as n increases, and the units are the square of the response units (what do we make of squared years?). One way to handle the dependence on n is to divide by n, thereby yielding the mean-squared error (MSE) metric:

$$MSE := \frac{1}{n} SSE$$

Then to ensure the error units match the response units, we can take the squared root, yielding the root-mean squared error (RMSE) metric:

$$RMSE := \sqrt{MSE}$$

Thus, to compare the performance of two models, we can compute their $RMSE$ on the same set of data points. We say the model with a smaller $RMSE$ performs better.

Another error popular error metric based on the SSE is the R^2 metric:

$$R^2 = 1 - \frac{SSE}{SST}$$

where SST is the SSE of what is known as the null model (discussed in a later section). The R^2 for a good model would ideally be close to 1, and closer to (or below) 0 would indicate poor performance. However, the R^2 metric is not as interpretable as $RMSE$. To provide users of this model with a simpler interpretation, I would use $RMSE$ instead of R^2.

9.1 In-Sample vs. Out-of-Sample Metrics

The discussion so far has mentioned computing the SSE for a collection of n points, but not which n points. When the collection of points on which the SSE is computed is precisely $\mathbb{D}$, the collection of points on which the model was trained, we call the corresponding metrics in-sample metrics. However, if we compute the error metrics on a set $\mathbb{D}_*$ (from the same phenomenon of course) but disjoint from $\mathbb{D}$, then we call the metrics out-of-sample metrics.

The distinction between in-sample and out-of-sample metrics is critical. It can be shown that in-sample metrics are dishonest, in the sense that it is possible to obtain excellent in-sample performance (all the way to an SSE of 0) by adding garbage features that have nothing to do with the phenomenon. Yet, the out-of-sample error metrics would increase, as they should in the face of unrelated information. This phenomenon is called overfitting, and it occurs because a larger set of features enables a larger set of candidate functions $\mathcal{H}$ that can be better tailored to the particular set $\mathbb{D}$. Put another way, overfitting is akin to getting lost in the noise while looking for a signal [Sil12]. In contrast, underfitting occurs when $\mathcal{H}$ is relatively small, whereby the model performs poorly because the functions are too simple to capture the functional relationship between the inputs and outputs. To detect overfitting, we need to compute out-of-sample metrics, which unlike in-sample metrics are honest. Another reason why the distinction matters is because the point of building a prediction model is to predict for out-of-sample data. For example, Luis from our earlier example is an out-of-sample data point, who had the fair question about how accurate our model is, which we can answer in terms of the out-of-sample $RMSE$.

It is likely that our approach to predicting human lifespan with 6 features and a linear model will severely underfit. We could address this by adding other features, such as the country a person lives, or how much money they earn, and choosing a more complicated set $\mathcal{H}$. However, as we discussed earlier, collecting data about the given 6 features we mentioned is already difficult, so data collection can be even more complicated if we add other predictors. Picking other candidate sets can also be difficult, especially when there is more than 1 input, since it is harder to visualize the data.

10 The Null Model

A simple idea we have not considered is whether we need any features or an algorithm at all. In other words, suppose we go on Wikipedia and learned the lifespan $y_1, y_2, \ldots, y_n$ of n different people. However, their Wikipedia page did not have information about the number of steps they walked each day, the number of hours they slept each night, or any of the features $x_1, x_2, x_3, x_4, x_5, x_6$ that we are using as proxies to the true drivers. In the absence of feature, there is still one model we can employ: the null model. Denoted g_0, the null model also comes from the hypothesis set $\mathcal{H}$, so its form depends on the functional forms allowed by $\mathcal{H}$. Thus, we know at the very least that g_0 gets its input from $\mathcal{X}$ and outputs a value in $\mathcal{Y}$. Since $\mathcal{Y} = \mathbb{R}$ is numeric, the null model corresponds to computing the arithmetic mean:

$$g_0(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} y_i$$

That is, to make a prediction about the lifespan of an individual whose features are $\boldsymbol{x}$, we actually ignore the features, and predict that their lifespan is simply the average of the lifespan of the n individuals we have seen. In some way, this is justified if we truly know nothing about each person other than how long they lived. The null model is useful because it can act as a reference for performance. It stands to reason that if we have accurate features about the health habits of an individual, if those features have any predictive value for lifespan, and if our algorithm is any good, then our computed model g ought to do better than g_0. To determine how well g_0 performs, we can compute its SSE, traditionally denoted as SSE_0 or equivalently SST for sum-of-squared totals. Thus, in addition to serving as a reasonable model when we don't have any features to make predictions, the performance for g_0 can serve as as threshold for performance of any other model we make.

11 Ordinary Least Squares Regression

To understand ordinary least squares, suppose that the number of features in our model is $p = 1$. For concreteness, say that we are only tracking x_1, the number of steps a person walks each day, hereby referred to as x for this example. Then our data set $\mathbb{D}$ is a list of points in the two-dimensional Cartesian coordinate plane:

$$\mathbb{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$$

Note that x_1 in this example is the number of steps the first person walks each day, not the generic name for the walk feature, and x_2 is the number of steps the second person sleeps, not the generic name for the daily sleep feature. We can visualize $\mathbb{D}$ using a scatterplot. In the past, you may have learned about the idea of the line of best fit. A general line has the equation $\hat{y} = w_0 + w_1 x$ for some parameters $w_0, w_1 \in \mathbb{R}$. For the value of the feature x_k corresponding to the number of steps the kth person takes daily, the line would predict $\hat{y}_k = w_0 + w_1 x_k$. The actual response for the kth person is y_k, so the error in the prediction with this line is $e_k = y_k - \hat{y}_k$. Since we have n points, that implies n residual errors $e_1, e_2, \dots, e_n$. The idea of ordinary least squares is to pick a pair of parameters $b_0, b_1 \in \mathbb{R}$ defining a line $\hat{y} = g(x) = b_0 + b_1 x$ so that the SSE given by

$$SSE := \sum_{k=1}^{n} e_k^2 = \sum_{k=1}^{n} (y_k - \hat{y}_k)^2$$

is as small as possible. Implicit is our discussion is the assumption that out algorithm $\mathcal{A}$ is searching through a space $\mathcal{H}$ of all possible lines, which is indeed the case since earlier de fined $\mathcal{H}$ to be the set of hyperplanes. The algorithm, then can be expressed as

$$\boldsymbol{b} = \underset{\boldsymbol{w} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \{SSE\}$$

where $\boldsymbol{b} = \begin{bmatrix} b_0 & b_1 & b_2 & \cdots & b_p \end{bmatrix}^\top$ is a vector of all the least square parameters. In our example, $p = 1$, so we would have $\boldsymbol{b} = \begin{bmatrix} b_0 & b_1 \end{bmatrix}^\top$. Note the procedure for the algorithm can be read as "find the parameters that minimize the SSE".

Admittedly, ordinary least squares may not be a suitable algorithm to use if our data set $\mathbb{D}$ does not suggest a linear trend. However, what is attractive is that under mild conditions it admits an exact solution $g = \mathcal{A}(\mathbb{D}, \mathcal{H})$. Indeed, throughout our discussion, we have had to give up hope on finding an exact solution for h^*, f, and t, so why would g be any different? It turns out that many minimization problems like the one for SSE do not have exact solutions and required complicated optimization algorithms to give an approximate solution. Thus, when it comes to being able to state a precise solution, OLS is the exception, not the rule.

That is not to say that g, the function determine by the OLS procedure, will be effective for making predictions. We know, for example, that there are some optimal weights $\beta_0, \beta_1 \in \mathbb{R}$ defining a linear function $h^*(x) = \beta_0 + \beta_1 x$ that our algorithm will not produce. Since our g is different from h^*, we accrue estimation error, given by

$$h^* - g = e - \epsilon$$

Moreover, if the lifespan phenomenon is not linear in the walking feature (say it's quadratic), then we incur misspecification error, given by

$$\epsilon - \delta = f - h$$

Of course, there is also ignorance error, which we cannot do anything about. An advantage of a linear model is that it is highly interpretable. For example, suppose we get

$$\hat{y} = g(x) = 20 + \frac{1}{1000}x$$

Then we can say that for every 1000 steps extra that a person takes daily on average, their lifespan increases by 1 year. Extending the OLS algorithm to more features is not much more difficult after introducing some notions of linear algebra and calculus.

12 Alternative Algorithms and Models

In an earlier section I made the decision to use the OLS algorithm to model the lifespan phenomenon. A merit of OLS is that it is heavily studied, so more precise statements can be made about its properties. However, it is worth considering what alternatives exist. For modeling a numeric response like lifespan, for example, we could also use a tree model with a regression tree partitioning algorithm. More generally, we may have M pairs of algorithms and models

$$(\mathcal{A}_1, \mathcal{H}_1), \ldots, (\mathcal{A}_M, \mathcal{H}_M)$$

How do we decide on which one is best? One approach is to take the training data $\mathbb{D}$ and split it into two pieces: $\mathbb{D}_{\text{train}}$ and $\mathbb{D}_{\text{test}}$. For each pair, we compute a prediction function on $\mathbb{D}_{\text{train}}$, and we compute an out-of-sample metric by predicting $\mathbb{D}_{\text{test}}$:

$$\begin{aligned} g_1 = \mathcal{A}_1(\mathbb{D}_{\text{train}}, \mathcal{H}_1) \quad &\implies g_1(\mathbb{D}_{\text{test}}) \quad \implies RMSE_1 \\ &\vdots \\ g_M = \mathcal{A}_M(\mathbb{D}_{\text{train}}, \mathcal{H}_M) \quad &\implies g_M\mathbb{D}_{\text{test}}) \quad \implies RMSE_M \end{aligned}$$

We would then choose the model $(\mathcal{A}_{m_*}, \mathcal{H}_{m_*}$ for which g_m* had the lowest $RMSE$. That is, model m_* is what we would deem to be our best model, so we can compute a final prediction function by training it not just on $\mathbb{D}_{\text{train}}$ but on all of $\mathbb{D}$:

$$g_{\text{final}} = \mathcal{A}_{m_*}(\mathbb{D}, \mathcal{H}_{m_*})$$

This basic approach has some limitations, but a variation of this idea can address some of them.

13 Conclusion

I believe the title of my essay is incorrect. The problem with a linear model is that it is too simple to describe the relationships between the factors that contribute to healthy aging. For example, the coefficients of each feature in a linear model have either a positive or negative sign, indicating that the feature always has a positive or a negative effect in the person's lifespan. However, it may be that a moderate amount of walking has positive effects, whereas walking 40,000 steps each day may have an adverse effect. A linear model will not capture this dynamic, and will suffer heavily from extrapolation.

References

[Bla21] Mikhail V Blagosklonny. “No limit to maximal lifespan in humans: how to beat a 122-year-old record”. In: *Oncoscience* 8 (Dec. 2021).

[BS16] Dan Buettner and Sam Skemp. “Blue Zones: Lessons from the world’s longest lived”. In: *Am. J. Lifestyle Med.* 10.5 (Sept. 2016), pp. 318–321.

[BSC16] Benjamin Baddeley, Sangeetha Sornalingam, and Max Cooper. “Sitting is the new smoking: where do we stand?” In: *Br. J. Gen. Pract.* 66.646 (May 2016), p. 258.

[Cen24] United States Census. *Population*. 2024. URL: https://www.census.gov/topics/population.html (visited on 04/02/2025).

[Dat+23] Saloni Dattani et al. “Life Expectancy”. In: *Our World in Data* (2023). https://ourworldindata.org/life-expectancy.

[Hic18] Tianna Hicklin. 2018. URL: https://www.nih.gov/news-events/nih-research-matters/healthy-habits-can-lengthen-life (visited on 03/31/2025).

[Ins25] Clay Mathematics Institute. *Navier-Stokes Equation*. 2025. URL: https://www.claymath.org/millennium/navier-stokes-equation/ (visited on 04/01/2025).

[Jae+12] Andreas Jaehne et al. “How smoking affects sleep: A polysomnographical analysis”. In: *Sleep Medicine* 13.10 (2012), pp. 1286–1292. ISSN: 1389-9457. DOI: https://doi.org/10.1016/j.sleep.2012.06.026. URL: https://www.sciencedirect.com/science/article/pii/S1389945712002882.

[Sil12] N. Silver. *The Signal and the Noise: Why So Many Predictions Fail-but Some Don’t*. Penguin Publishing Group, 2012. ISBN: 9781101595954. URL: https://books.google.com/books?id=SI-VqAT4_hYC.