36-402 Project 1

Edwin Baik (ebaik)

4/2/2021

Introduction

Preston Jorgensen, a billionare, is interested in finding the key to immortality, and has asked me to find variables that can help his search for immortality. We must first understand the relationship between variables given within the AnAge data set and the resulting lifespan, and search for correlations that can potentially help a human become immortal. Variables like metabolic rate can be critical in understanding what can increase lifespan. Furthermore, we wish to find the lifespan for a crab-eating raccoon with 50% less metabolic rate. If there is a proper relationship between lowering the metabolic rate to incrase one's lifespan, then Mr. Jorgensen is willing to put in treatment for the crab-eating raccoon to increase its lifespan **(1)** . We are using data provided from the AnAge database. There are 347 observations of 14 variables, with 7 relevant variables being the HAGRID (Human Ageing Genomics Resources ID), Kingdom, Common Name, maximum.longevity.yrs (lifespan), body.mass.g (body mass), Metabolic rate, Temperature **(2)**. From our final findings and the model we formulated, it is fair to say that metabolic rate CAN change the maximum longevity of years one can live. While there are alot of variables ranging from Temperature to mass that can affect the maximum longevity of years, metabolic rate is one variable that can also change the lifespan. Specifically, if one would reduce their metabolic rate, then your lifespan can increase in return. For example, given the crab-eating raccoon, it is possible to reduce its metabolic rate by 50% and have its expected maximum lifespan be approximately between 19.13 and 23.41 years **(3)**.

Exploratory Data Analysis

We first create a variable called Metabolic.by.mass, which represents the amount of energy used per unit of body mass. We derive this variable by dividing the Metabolic rate by the body mass of the animal **(1)**.

Using the data provided by AnAge database, and the 347 observations within the data, we will use a portion of the variables given to answer the next few questions. Note that we first define 4 key variables. The first will be our response variable, which is denoted as maximum.longevity.yrs, or also known as the maximum lifespan of an animal, in years. Our key explanatory variable will be the Metabolic.by.mass, and as described before, is the amount of energy used (metabolic rate) per unit of body mass. Another explanatory variable we can potentially use is Temperature, and it is measured in Kelvin. Finally, we have a control variable defined as the Class variable, and is separated by the Amphibia, Aves, Mammalia, and Reptilia Classes. The data set did include other variables like Kingdom, Phylum, Order, Family, Genus, Species, and HAGRID, but we did not believe these variables were necessary as explanatory or confounding variables in our analysis. For example, we noted that Phylum and Kingdom were the same for all the dataset examples, so no distinguishment could occur between the dataset examples. As noted before, we use a database provided from the AnAge space, and it is assumed that the data has already been cleaned of missing values from variables. The first of the variables is our response variable, maximum.longevity.yrs (lifespan) within the data. Note its distribution is extremely right-skewed overall, as shown by its histogram, as shown below in Figure 1. This indicates that the median of the lifespan data is less than the mean of the data. Furthermore, there is low variability of the lifespan data, as indicated by the non-distributed values across the lifespan values. The heavy right-skew indicates a large amount of outliers, and also raises the idea that a transformation should be used on the response variable, potentially a logarithmic transformation **(3)**. When we then look at the explanatory variables, Temperature and Metabolic Rate by Mass, and their normality plots, a few things can be gleaned, as found in Figure 2. Note that the temperature is extremely left-skewed, with only a few outliers towards the left/middle ends of the graph. Furthermore, the other explanatory variable, the Metabolic Rate by Mass, also is very right-skewed, indicating yet another potential, but necessary transformation to normalize the data. We also took a summary of the categorical variable Class, and recognized that there are 315 Mammalians, 14 Reptilians, 2 Avians, and 16 Amphibians. From this, we can glean that the Aves class may not be of much use in creating a conclusion as there are only 2 animals with the Aves class **(2)**.

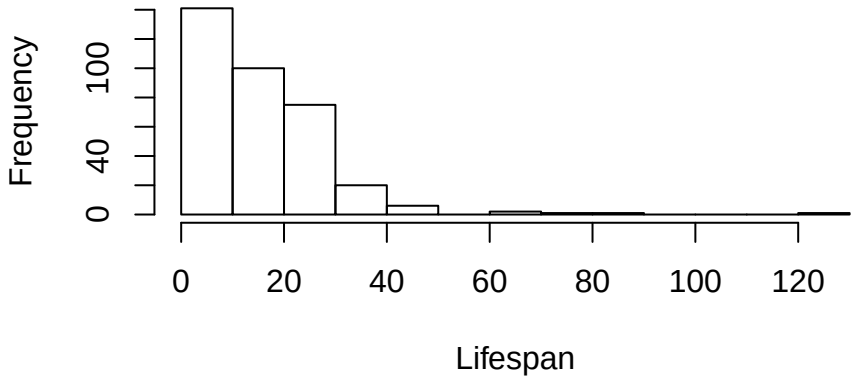

Figure 1: Histogram Regarding Lifespan Distribution

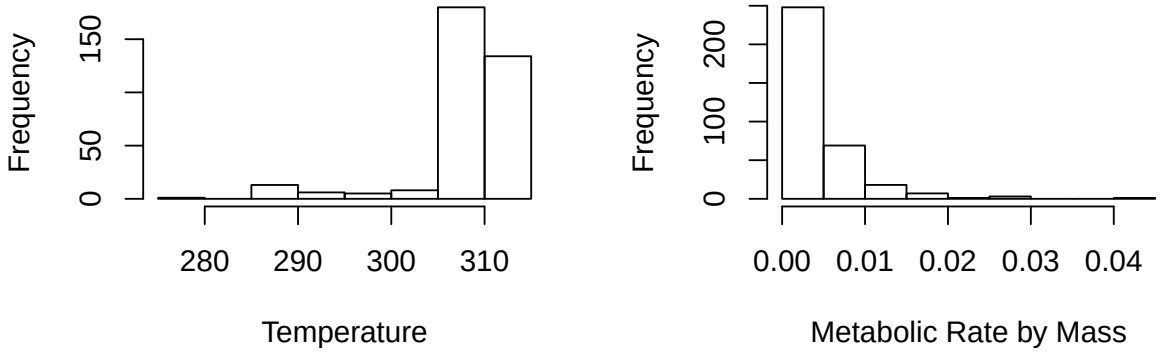

Figure 2: Histograms of Explanatory Variables

Next, we wished to see how the variables that we chose were related to each other so that when we do modeling, we can create a better idea of how to fit models to the data. When we compare our predictors to our response variable, we note a few ideas. The correlation between temperature and lifespan was -0.221, while the correlation between metabolic rate by mass and lifespan was 0.385. This indicates that there is some negative relationship between the two predictor variables and the response, whilst weak. The scatterplot matrix also helps us to find potential collinearlty in between the explanatory variables, Note that Between the two explanatory variables, there is a correlation of 0.260, and as shown below in Figure 3, there is some pattern of the points, as they are clustered at the top left of the scatterplot. This indicates some relationship between the two, and we can potentially need an interaction when we model to account for this relationship. Given this and our univariate EDA, as mentioned before we would like to do multiple transformations on our data. We would like to do a log transformation on BOTH the response variable lifespan and the explanatory variables Metabolic.by.Mass and Temperature. After the transformations, we recognize that the relationships seem to be much more linear, which may be accredited to the fact that much of the data has become normalized rather than the left or right skews that were produced initially. **(4)**.

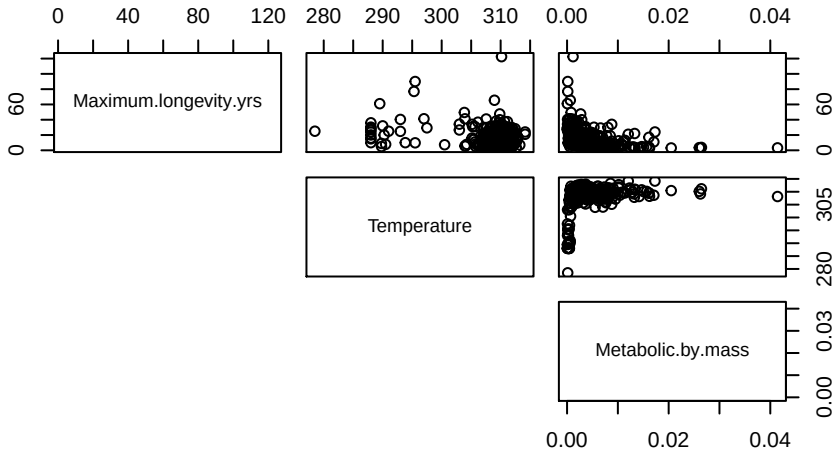

Figure 3: Scatterplot Matrix of Response/Explanatory Variables

For plots that can help us answer the research question, we can create a scatterplot between the lifespan and metabolic rate by mass both pre and post transformation. As recognized in Figure 4, the plot on the left indicates the relationship pre-logarithmic transformation, and shows a massive cluster of points towards the bottom left. However, the plot

on the right shows a much more negative relationship overall between the two variables, as indicated by the downward trend of points from the left to right. **(5)**.

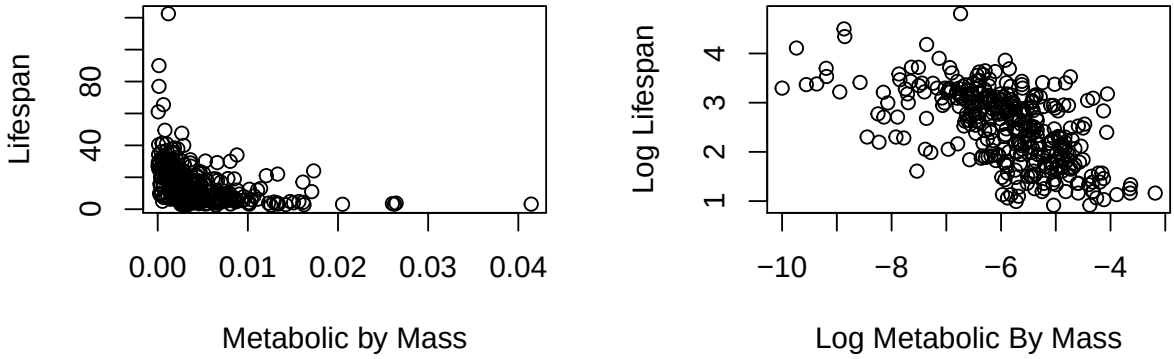

Figure 4: Scatterplots Between Lifespan and Metabolic Rate by Mass

Some trends or interesting features that I may see within the analysis is that due to the negative relationship between the metabolic by mass and the lifespan, there is clearly going to be a negative return on the lifespan as your metabolic by mass decreases. This indicates that if one's animal mass decreases but its metabolic rate stays the same, the lifespan can potentially increase. Inversely, if an animal's metabolic rate increases but its mass stays the same, the lifespan may decrease **(6)**!

Modeling & Diagnostics

We can first construct a linear model to predict the maximum.longevity.yrs (Lifespan) using the explanatory variable Metabolic.by.mass, but also use the explanatory variable Temperature. Note also that the information we pass for our linear model will be the log transformed versions of both the explanatory variables and the response variable because we indicated a more normal distribution of the two after the transformations applies. Furthermore, we do NOT include an interaction term as the interaction between the two explanatory variables are NOT significant via an ANOVA test. The reason we chose the variables we did for our linear model (transformed Temperature and transformed Metabolic.by.mass) was through process of elimination. The two continuous variables that we could've used was initially both Temp and Metabolic.by.mass. When we formed

the model, we recognized that both variables were significant (Temperature -> 5.53e-09, Metabolic.by.mass -> 2e-16). Note that we used the transformed values because the initial EDA showed a non-relationship between the non-transformed variables. We wished to see if we could include or not include the log-transformed temperature into our model, so we ran a F-test. The F-test came as significant, indicating that the temperature term was necessary in the prediction for our log-transformed lifespan **(1)**.

Linear: $log(Lifespan) = B_0 + B_1 * log(MetabolicByMass) + B_2 * log(Temperature)$

Since the linear model does not seem sufficient enough, we fit multiple smoothing splines upon the same data with the respective degrees of freedom of 3, 4, 5, 6, and 7. Note here that we created smoothing spline models that only had the log metabolic.by.mass be our only predictor for the log lifespan **(2)**.

We used 5-fold cross validation to estimate the prediction errors of the models that we created. Cross-validation is required because it helps to choose a model that is best fit for the data as the diagnostics and goodness of fit from a model is not enough to make that choice. After using 5-fold cross-validation over the linear model, and the smooth splined models that varied from 3-7 degrees of freedom, we recognized that the linear model between log lifespan vs. log metabolic.by.mass and log temperature had the lowest prediction error of them all **(3)**. The model errors are shown below:

a) (Linear Model) MSE: 0.3340750, SE: 0.02845663
b) (Spline DF = 3) MSE: 0.3594142, SE: 0.02610298
c) (Spline DF = 4) MSE: 0.3581587, SE: 0.02660235
d) (Spline DF = 5) MSE: 0.3577763, SE: 0.02685368
e) (Spline DF = 6) MSE: 0.3583246, SE: 0.02720832
f) (Spline DF = 7) MSE: 0.3598446, SE: 0.02766100

When we look at our chosen model, the linear model with the two explanatory variables, some information can be gleaned from the diagnostics and residual plots. Note that within the residual plots of Figure 5, the residuals are well distributed both above and below the residual line and there are no patterns being formed. This helps us to assume linearity of the dataset. Furthermore, the spread of the residuals seem proper, indicating constant variance for our model. However, there are signs of clustering within the temperature residuals, which could cause some violations in linearity. To combat this potential violation, we can try to spread out the temperature even more from its normal distribution through further transformation. We should be careful to do more transformations as too many can potentially cause overfitting **(4)**. Unfortunately, the difference

between the models regarding their prediction error does NOT seem significant enough to warrant a definitive choosing of the model. Note that there seems to be a large difference in the MSE between our chosen model and the rest of the smoothign spline models, creating confidence that that the difference seems significant **(5)**. Given that our data was transformed to become normally distributed, and the residuals show no signs of violation regarding the normality and constant variance assumptions, we can safely say that we can use the parametric bootstrap on our model. Note we would only use the non-parametric bootstrap if we could not assume the normality of our dataset **(6)**.

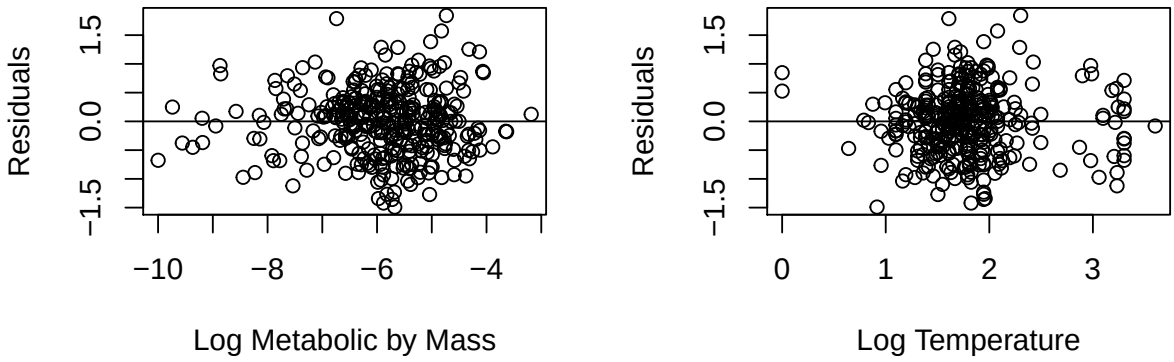

Figure 5: Residual plots of the linear model

Results

Given the linear model that we chose, there are multiple ways to determine whether animals with lower metabolic rates have longer lifespans. When we run an F-test on our model, this indicates if there is some relationship between the predictor and the response variables. Since the F-test provides a very large F-statistic, this in turn gives us an extremely small p-value that is less than 2.2eˆ-16. However, our model contains both the log metabolic.by.mass AND the temperature, so an F-test is not enough to guarantee that lower metabolic rates have longer lifespans. Rather it provides us with an initial overview. When we run a t-test between the predictor log maximum.longevity.yrs and the log metabolic by mass, we find that the p-value is again less than 2eˆ-16. This indicates a significant relationship between the two values. Furthermore, since our coefficient

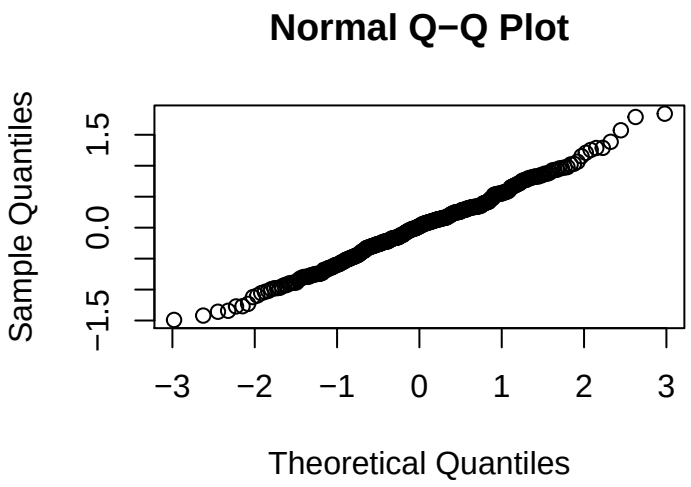

Figure 6: QQ Plot of the residuals for the linear model

of our model regarding log metabolic.by.mass is -0.51621 approximately, this indicates the more metabolic.by.mass decreases, the more lifespan increases. This means that a lower metabolic rate under the assumption that the mass is not proportionally changing to the lower metabolic rate CAN lead to a potential increase in the maximum longevity in years **(1)**.

As noted, Mr. Jorgensen wished for us to find the crab-eating raccoon and its features, and measure its lifespan given that its metabolic rate was 50% smaller. We first take the features of the animal at HAGRID value 1898 (value of the crab-eating raccoon). We then take its metabolic rate, decrease it by 50%, and then divide it over its mass. We then find the log value of that metabolic rate per mass, and also take the log value of the temperature of the animal. Using these two new explanatory variables, we can predict the log lifespan. Noting that the regular maximum.longevity.yrs for a crab-eating raccoon is 19 years, with our new prediction given our linear model, we recognize that the maximum.longevity.yrs given 50% less metabolic rate is approximately 21.13 years **(2)**.

When we did bootstrapping, we used a parametric approach due to our satisfaction of assumptions given our final model. We parametrically bootstrapped a new model by resampling on the residuals, and then took the 1000 predictions created from our 1000 models to create a confidence interval. Ultimately, we had a confidence interval of 19.13 to 23.41, indicating that we have 95% confidence the max.longevity.yrs of the crab-eating raccoon will be approximately between these two values **(3)**.

Conclusions

Given our findings regarding the relationship between metabolic rate and lifespan, there are multiple ideas we can glean. First, using our t-test, we can assume that there is some significant relationship between metabolic rate and lifespan. Furthermore, given the model coefficients from our best chosen model, we can argue that there is an inverse relationship between the two variables. In translation, the lower the metabolic rate, the higher the lifespan can potentially be **(1)**. We cannot definitively say that reducing the crab-eating raccoon's metabolic rate by 50% would CAUSE its lifespan to change. Whilst we did a t-test between metabolic.by.mass and the lifespan, association does not equal causation, and this is also an observational study, further not allowing us to assume causality. However, we can say that by our linear model, decreasing the metabolic rate would change the maximum longevity in years. In our Results, we recognized that if the crab-eating raccoon's metabolic rate was decreased by 50%, it would increase the maximum longevity of years from the usual 19 years to an approximate 21.13 years. This indicates a 10% increase in the lifespan, which is a large difference between the years lived **(2)**. One of the first limitations to our data was having to use logarithmic transformations in order to normalize our data. Transformations can be a cause in overfitting of the data, and whilst necessary to do transformations to have a more normal distribution, our model does rely on them for non-violations in our assumptions. Another limitation is the ability to categorize between different animals/species. Currently, our model can only predict the maximum longevity of years only given the metabolic rate by mass and temperature, and does not work well with trying to categorize the lifespan given a category of animals. For example, it is difficult to find the maximum longevity of years given some metabolic rate, mass, and temperature for the Mammalia class, or the Genus. This becomes a limitation as the model can only work on specific animals, and not for a broader understanding of the animal kingdom **(3)**.