36-402 DA Exam 1

Jacky Liu (jackyl1)

3/19/2021

Introduction

Preston Jorgenson is interested in determining what factors appear to be related to lifespan. Given in-depth information and data from hundreds of scientific papers for over 4,200 species, we will try to answer the following three questions. (1) The first is whether the slowing of metabolic rate increases lifespan. We will develop a model for lifespan using metabolic rate while trying to control other variables. The second is whether the relationship between metabolic rate and lifespan is nonlinear. We will use a nonparametric model to analyze this fit. Lastly, we will predict the mean lifespan of an animal whose metabolic rate is reduced by 50%. (2) The data we are using to fit these models, which is from AnAge for animals in the Chordata phylum, includes 347 observations and 14 variables.

Our analysis led us to the conclusion that there is a relationship between life expectancy and metabolic rate in animals (3) With certain assumptions, we predicted that decreasing the metabolic rate of an animal is likely to lead to an increase in life expactancy.

Exploratory Data Analysis

The data being used is from The AnAge Database of Animal Ageing and Longevity, which was last updated in 2017, to study the ageing and lifespans of various species of animals and answer our three questions. We define several key variables: The first is our response variable which we will call lifespan (2). This is the maximum longevity of the animal in years. For our explanatory variables, we first define body mass, which is the typical adult body mass of the animal in grams. Next, we define metabolic rate as the typical resting

metabolic rate, which is the rate in which energy is used, in watts. Finally, we define temperature as the typical body temperature of the animal in Kelvin units. In addition, we define a new variable called Metabolic By Mass which is the Metabolic rate divided by body mass in order to compare animals of different sizes. We will be using this variable instead of Metabolic rate. First, we will take a look at our response variable.

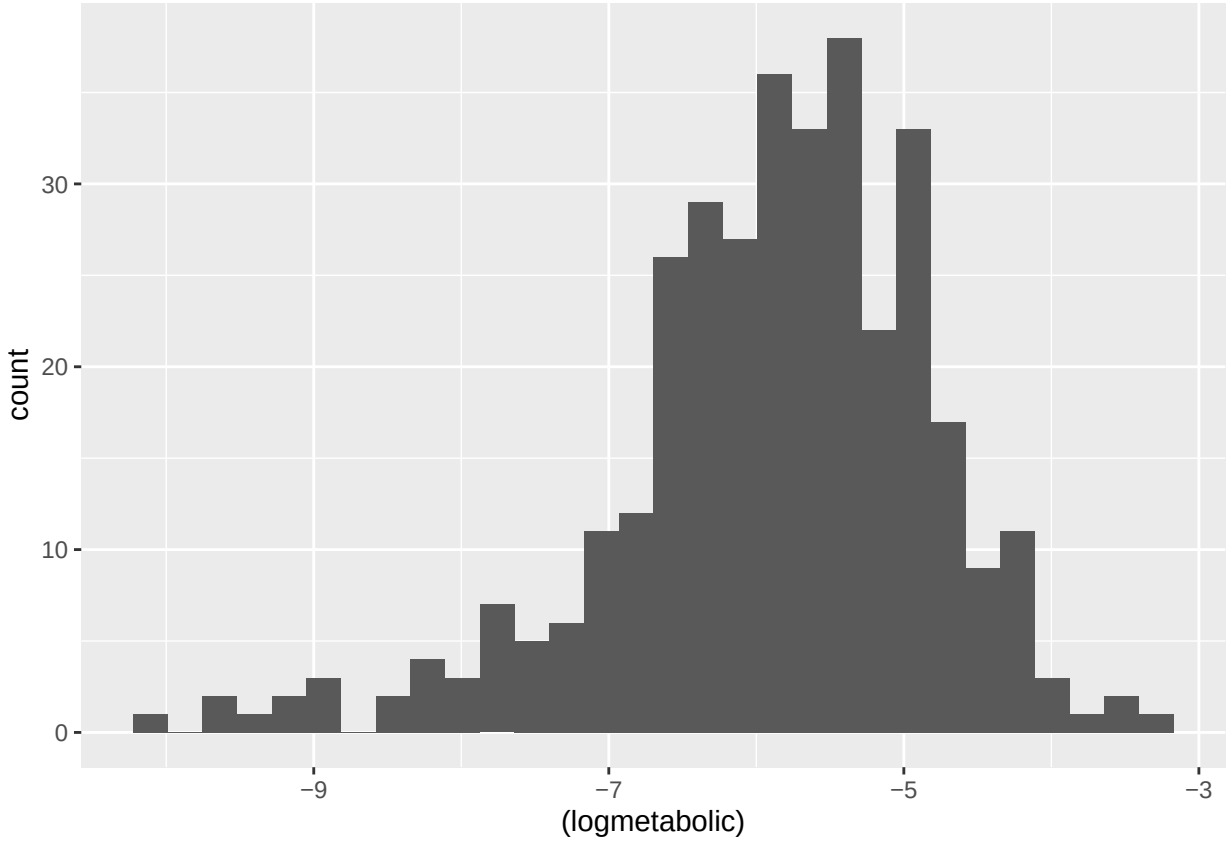


Figure 1: Histogram showing marginal distribution of an explantory variable, log metabolic by mass

Our response variable, which we defined as lifespan, has a very right-skew distribution (3). The majority of animals we are looking at have a life expectancy of under 25, with most animals having between approximately 3-10 years in lifespan. There are also clear outliers ranging from 50 years to 125 years. The mean lifespan is 16 years, median is 12.9 years, and the standard deviation is 12.8 years. This means that most, if not all of the animals we're studying will have significantly shorter lifespans compared to humans, so it is important to keep this in mind if the data is used in regards to human lifespans. (3) Due to this extreme skewness, we will log transform lifespan variable and will be using it for all of our models.Next, let's look at the most important explanatory variable for our

analysis purposes, which is metabolic rate. More specifically, Metabolic By Mass which we created.

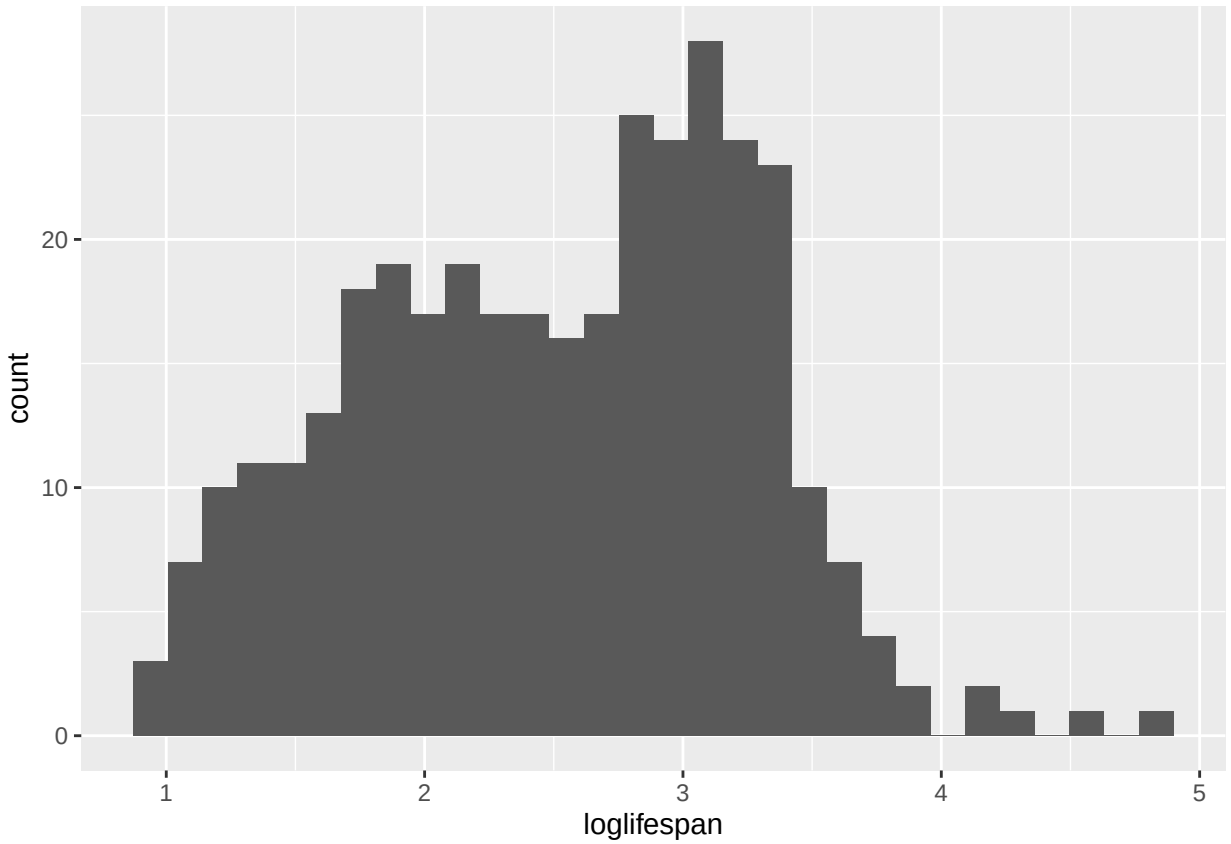
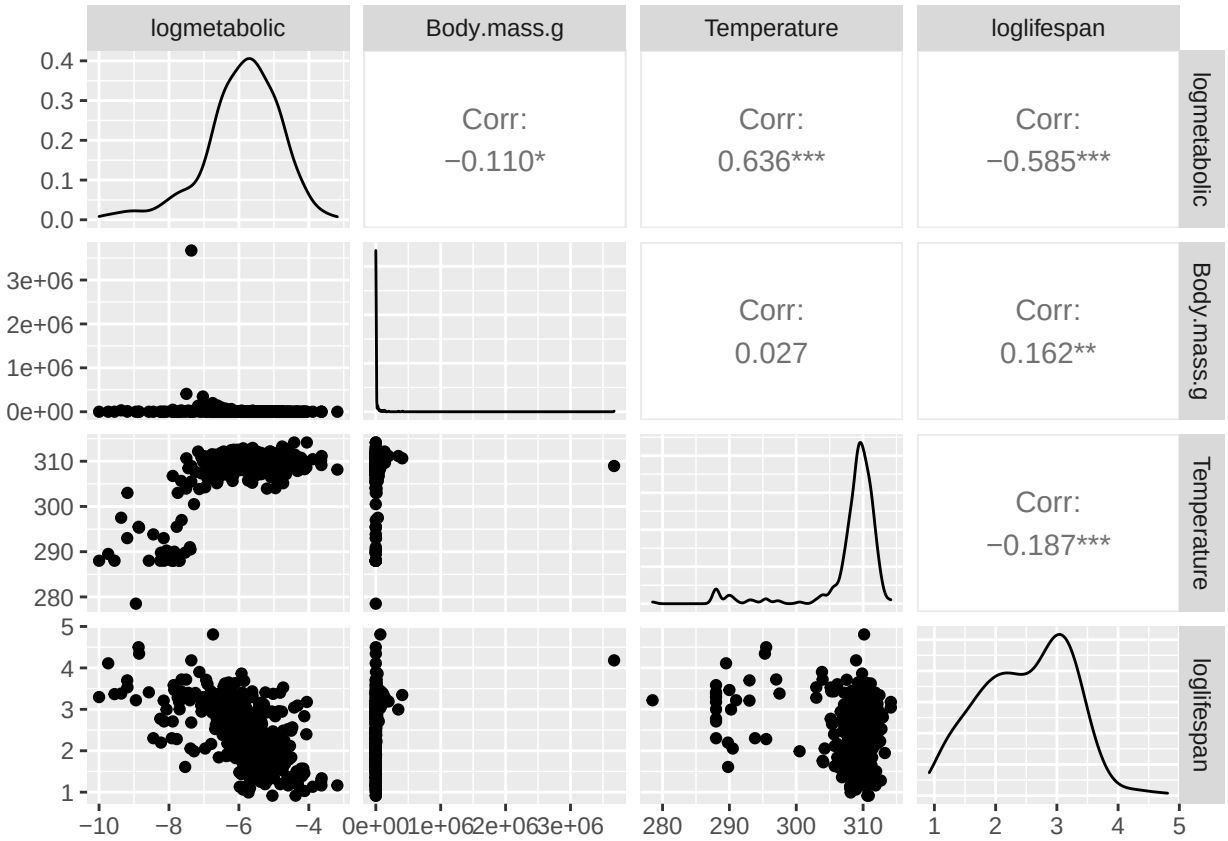


Figure 2: Histogram showing marginal distribution of the response variable, log lifespan.

Similar to our response variable, Metabolic By Mass also has a very similar right-skew marginal distribution. It wouldn't be surprising to find a correlation between the two. Metabolic by mass has a mean of 0.0044 watts per gram and a standard deviation of 0.0046 watts per gram. We will also log transform our metabolic by mass variable. Next, we would like to explore the relationship between the response and our predictors.



From here on, we will be referring to our log transformed Metabolic by mass variable as log metabolic and our log transformed Maximum life expectancy variable as log lifespan. From our multivariate exploratory data analysis, we can see a moderately strong negative correlation between log metabolic and log lifespan (4). There is also a notable strong positive correlation between temperature and log metabolic. This is interesting to see, but not surprising as studies have shown that a cooler core body temperature has been shown to slow metabolism. However it is irrelevant to our study because humans cannot lower their body temperatures to match that of animals (6). The correlation between log metabolic and log lifespan is important to note especially when we create our models because the primary goal of our study is to determine the relationship between metabolic rate and lifespan.

Modeling & Diagnostics

We constructed a linear model and a smoothing spline model in order to answer the questions Preston Jorgensen asked of us, which can be seen below (1).

Linear: lm(log(lifespan) ~ log(metabolic) + temperature)

Smooth Spline: smooth.spline(log(lifespan) ~ log(metabolic))

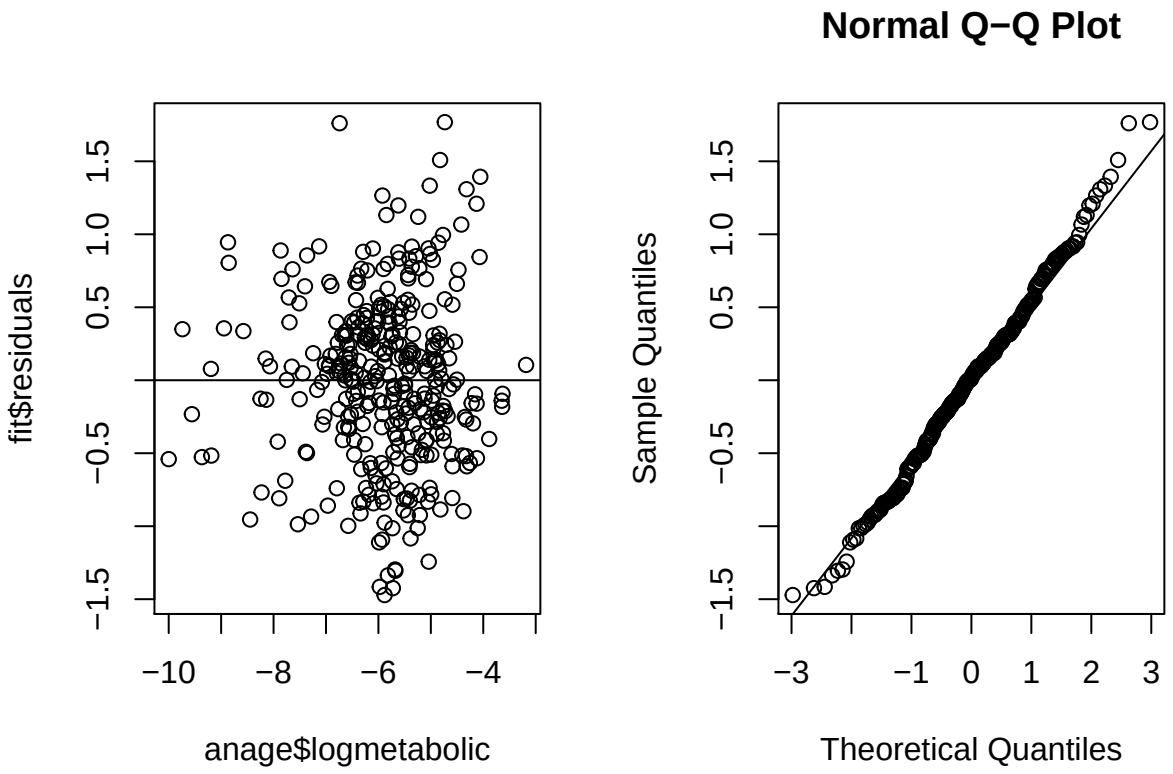


Figure 3: Diagnostic plots for our linear model

For our linear model, we started with all of our quantitative variables except for Metabolic rate because we have previously decided to use Metabolic by mass instead. Because body mass had a large p-value of 0.103 and was uncorrelated to anything in our 2D exploratory data analysis prior, we removed it. Temperature was kept as a confounding variable. Log transformations were performed on both our response variable, lifespan, and was performed on the explanatory varaible metabolic by mass.

The assumptions of the linear model (iid, linearity, constant variance, Gaussian noise) seem relatively but not perfectly plausible. Although the Q-Q line seems to fit the points very well, there may be signs of heteroskedasticity in the residuals plot. Overall the data show some modest signs of betraying the linear model assumptions, but not too extreme.

Next, we use a smoothing spline to fit the same data as in the linear model. Using 5-fold cross validation, we determined that between degrees of freedom 3 to 7, the model with

6 degrees of freedom has the lowest prediction error (2).

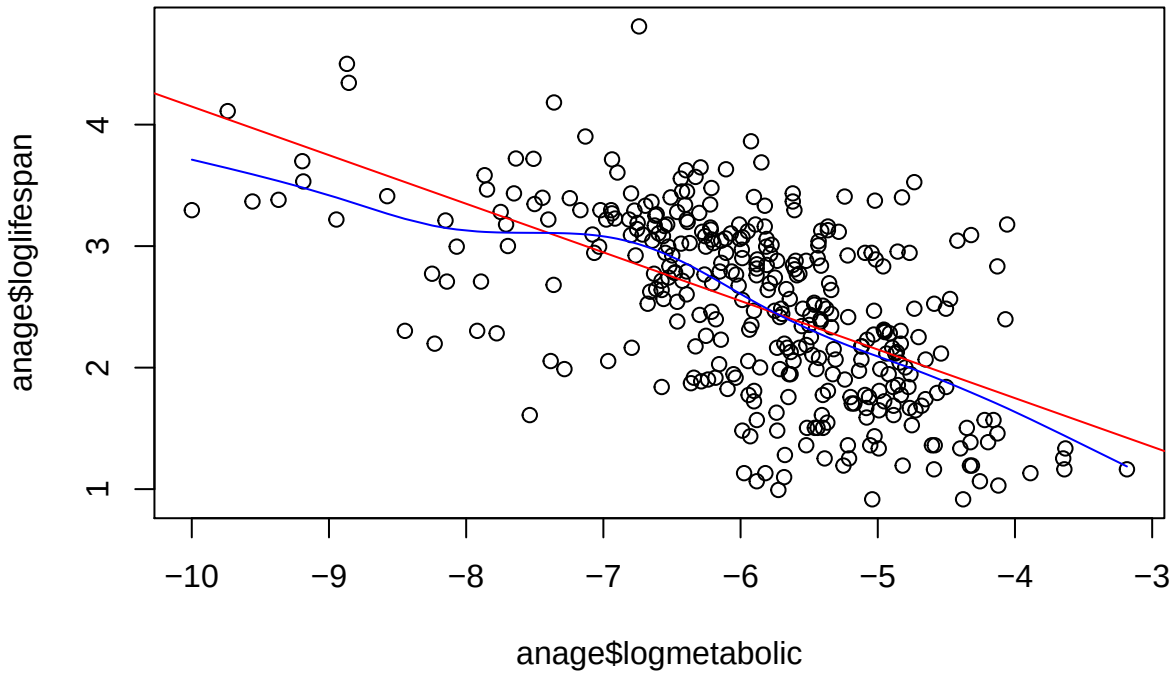


Figure 4: Both models plotted against our data. Red=linear model, Blue=smoothspline

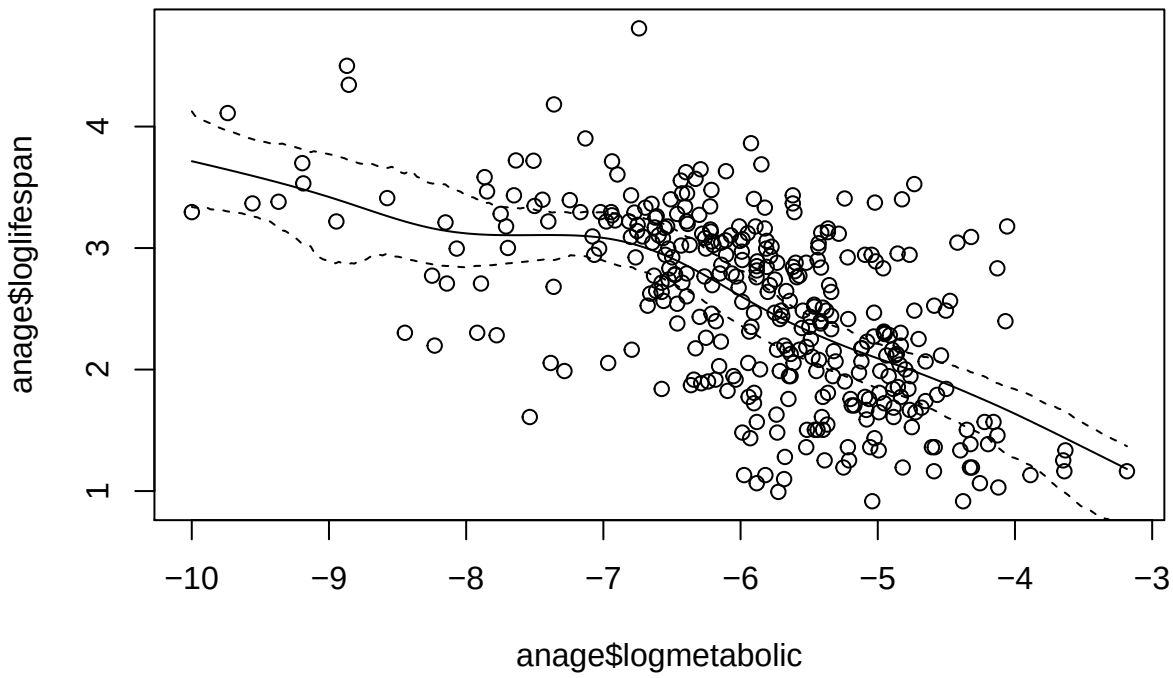
Results of CV	Linear Model	Smoothing Spline
Estimated MSE	0.4483198	0.3570683
Standard Error	0.03641642	0.02512945

Looking at figure 5, (4) we can see that both the linear model and smoothing spline model fit the data reasonably well. Upon closer inspection, when we compare the errors, we can see that the smoothing spline has a lower mean squared error (3). Therefore, we decided that the smoothing spline model is the best in this case and choose it to use to answer Preston Jorgenson's questions. Note that the difference between these two models is less than 0.01 which subjectively may be determined as significant or insignificant (5), but we choose the model with the lower error anyways for the purpose of our analysis.

Results

Using our smooth spline model to answer Preston Jorgenson's questions, we check the summary coefficients of our spline. With this nonlinear smoothing function, we determine that animals with lower metabolic rates indeed do have longer lifespans (p=0.003415) under a significance level of alpha=0.05 (1). In addition, looking at the graph of the smooth spline, we can see that log lifespan decreases steadily and drops off more significantly after log metabolic is greater than -6. Inversely, animals with less log metabolic rates have greater log lifespans. The negative values for log metabolic makes sense because most of our values for metabolic by mass are smaller than 1, and the log of a number between 0 and 1 noninclusive are negative.

First of all our model predicts that the log lifespan of a crab-eating raccoon is 2.675 years which corresponds to a lifespan of 14.51 years. Now, reducing the metabolic rate to 50% smaller we get a prediction of log lifespan equal to 3.033877 years, or 20.78 years. Although our model initally incorrectly predicted the crab-eating raccoon's lifespan with an error of approximately 4 years, if we compare the results of 100% metabolic rate vs 50% metabolic rate using our predictions, our model predicts that the lifespan of the crab-eating raccoon increases from approximately 15 years to 21 years, a 6 year increase in lifespan (2). Using bootstrapping with 1000 iterations, we get a confidence interval for log lifespan of crab-eating raccoon with 50% of its metabolic rate, which is (2.8599, 3.2719). This gives a nonlog 95% confidence interval of (17.46, 26.36) years for a crab-eating raccoon with 50% of its metabolic rate (3). Since the original predicted lifespan is not in our confidence interval for lifespan with 50% of metabolic rate, we can say that a crab-eating raccoon with 50% of its metabolic rate is indeed likely to live significantly longer.



.\

Conclusions

In this study, we used a smoothing spline model to examine the relationship between the life expectancy of animals and the metabolic rate of animals. We concluded that there was a significant relationship between log metabolic rate and log life expectancy (1). We predicted that a crab-eating raccoon, which our model shows has a life expectancy of 14.59 years from its normal metabolic rate, would expect to live between 17.46 years and 26.36 years if we decreased its metabolic rate by 50%. This interval is substantially higher than what a normal crab-eating raccoon would expect to live, so Preston Jorgenson can indeed conclude that reducing the crab-eating raccoon's metabolic rate by 50% would cause its lifespan to increase an average of 4 years (2). Looking at our analysis, there was not a super strong correlation between our two main variables, which made it somewhat difficult to fit a good model to the data. In addition, some of the assumptions of our model were not perfectly met, which may limit the credibility of our results. As we studied many different species of animals of different metabolism rates, we did not have same

species animals which had different metabolism rates, which makes it somewhat risky to jump to the conclusion of implied causality, as species of animals with lower metabolism having higher lifespans may just be a correlation that even our bootstrapped analysis cannot prove causality. (3) In addition, having a large size than 347 species and more variables that may be confounding (such as environment, diet, etc) can help us better understand the relationship between life expectancy and metabolism. Finally, because Preston Jorgensen is interested in extending his own lifespan, we must acknowledge the fact that even if there is a causal relationship between metabolic rates and lifespan in animals, this does not necessarily apply to humans.