36-402 DA Exam 1

Raymond (rli3)

3/19/2021

Introduction

For ages people have searched for the formula that granted immortality. There are endless possibilities in how this formula can be presented. While some think it is obtained through scientific research, others believe it can be found through a magical fountain. Eccentric billionaire, Preston Jorgensen, is another who is entranced in the race towards immortality. He believes the key to obtaining immortality is in performing statistical analysis on a data set that includes various information on thousands of different species of life. These pieces of information includes scientific classification, body mass, body temperature, and more.

With the access to this information, Jorgensen would like to know three things. **(1)** First, use a model for lifespan using metabolic rate to determine if lowering metabolic rate can increase lifespan. Next, use a nonparametric model to determine if a non-linear fit or a linear fit makes more sense when comparing lifespan and metabolism (even after transformations). Last, use a confidence interval to find if reducing the metabolic rate of his favorite animal (the crab-eating raccoon) by 50% will lead to a statistically meaningful increase to its lifespan.

(2) sourced used for this research is the AnAge Database of Animal Ageing, which is a database curated for the use of biological studies. More specifically, the database, which contains data on over 4200 species, is used to study ageing and lifespan.

The specific csv file we are using is a subset of that data set that contains 347 unique observations of 14 features. These features include *HAGRID* (a unique identifier for this entire data set), *Kingdom*, *Phylum*, *Class*, *Order*, *Family*, *Genus*, *Species* (which are the seven levels of scientific classification of life, in general to specific order), *Common.name* (the non-scientific identifier of the life form), *Maximum.longevity.yrs* (maximum lifespan, in years), *Body.mass.g* (standard adult body mass, in grams), *Metabolic.rate* (standard resting metabolic rate, in watts), and *Temperature* (standard body temperature, in Kelvin). The first feature, *X*, is most likely the row number of the data after extracting null-value rows- and is a functionally useless feature for our study.

(3) Through this research, it has been concluded that decreasing metabolism rate can increase lifespan. The relationship between the two is close to linear, though a non-linear model picked as the best possible model to compare the two. Lastly, the crab-eating raccoon’s lifespan is projected to statistically improve if its metabolism is cut in half, though not by much.

Exploratory Data Analysis

(1) A new variable will be created, called ‘Metabolic.by.mass’, which divides the metabolism feature by their respective mass values. This new variable represents the amount of energy used per unit of body mass for an animal, which allows comparisons between animals of different sizes.

Adding this new variable will allow us to compare metabolic rate to lifespan, controlling for body mass. This variable will replace ‘Metabolic.rate’ for the rest of the report.

(2) Looking through the distributions of our quantitative features (not shown), Temperature is skewed to the left with some lower bound outliers. All other numerical variables are skewed right with upper bound outliers. Since metabolism by mass is the only feature needed to answer the questions, I will show its distribution in Figure 1.

The distribution's high level of skewness indicates a potential transformation to be made. In Figure 2, the same distribution is shown, but with a log transformation done on the variable. The result looks much more normal (but still not perfect) and can later be used in the linear predictor.

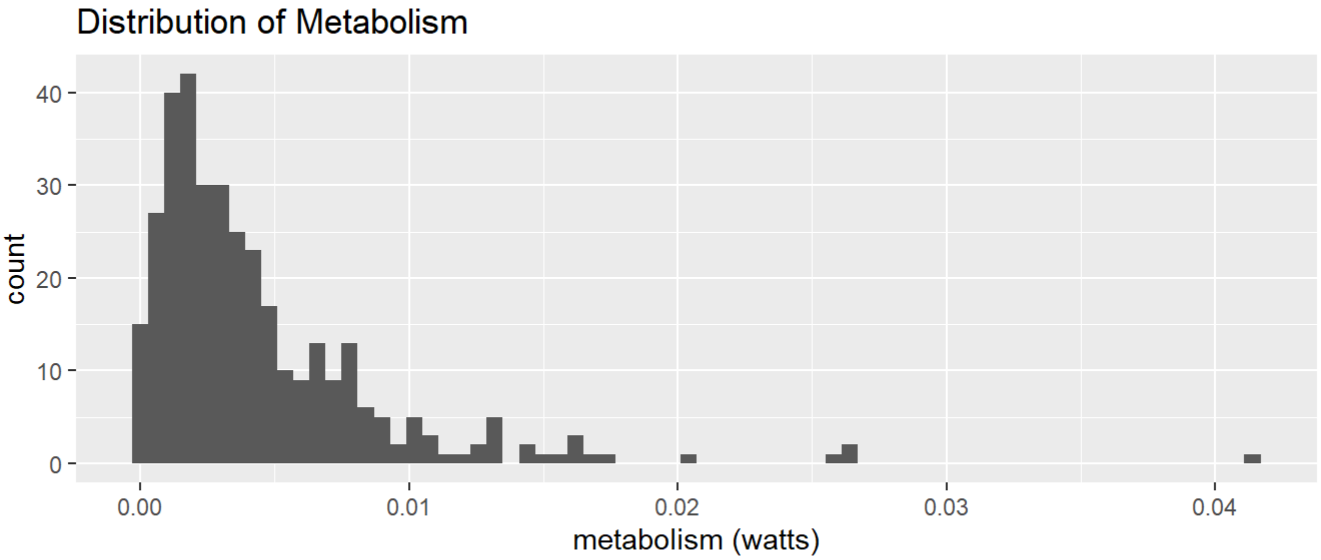

Figure 1: Distribution of animal metabloic rates by mass

.

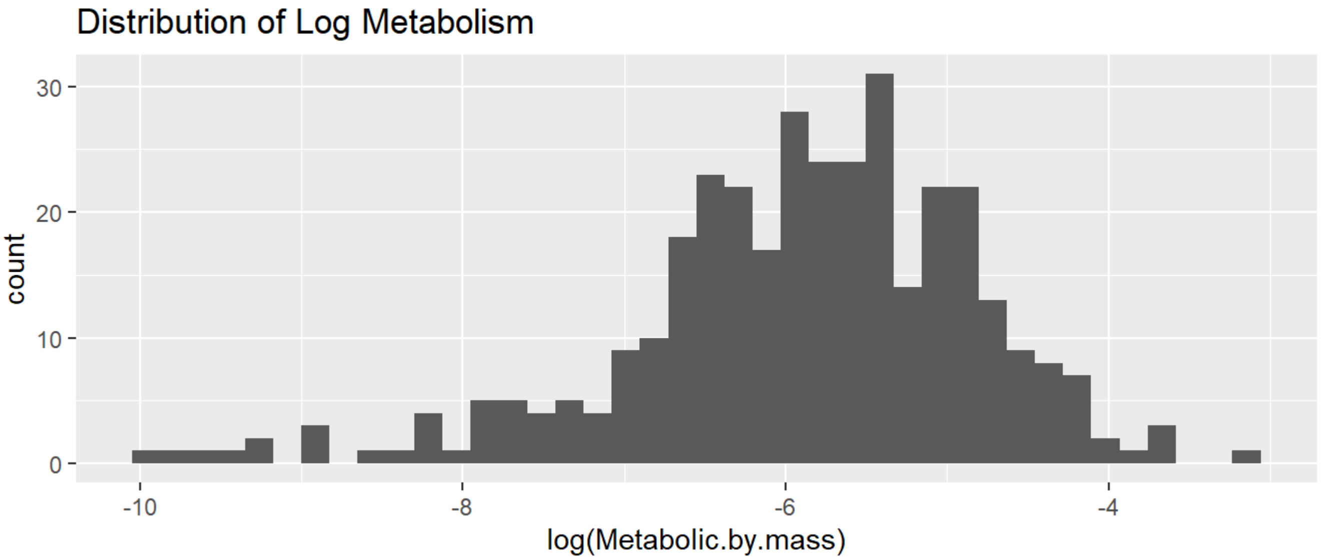

Figure 2: Distribution of log metabloic rates by mass

(3) The response variable is Maximum.longevity.yrs. This is the feature that represents lifespan, which directly reflects the tasks relayed in the introduction.

The distribution of the response variable is skewed right with the median lifespan at about 15-20 years. The vast majority of data points contain lifespans of below 40 years, but there are a number of upper outliers scattered across the histogram that goes up to 125 years. These findings are shown in Figure 3. Again, the skew is addressed by performing a log transformation. The new histogram, outlined as Figure 4, is still not perfectly normal. However, it is much better than the non-transformed distribution and can be used in our model going forward.

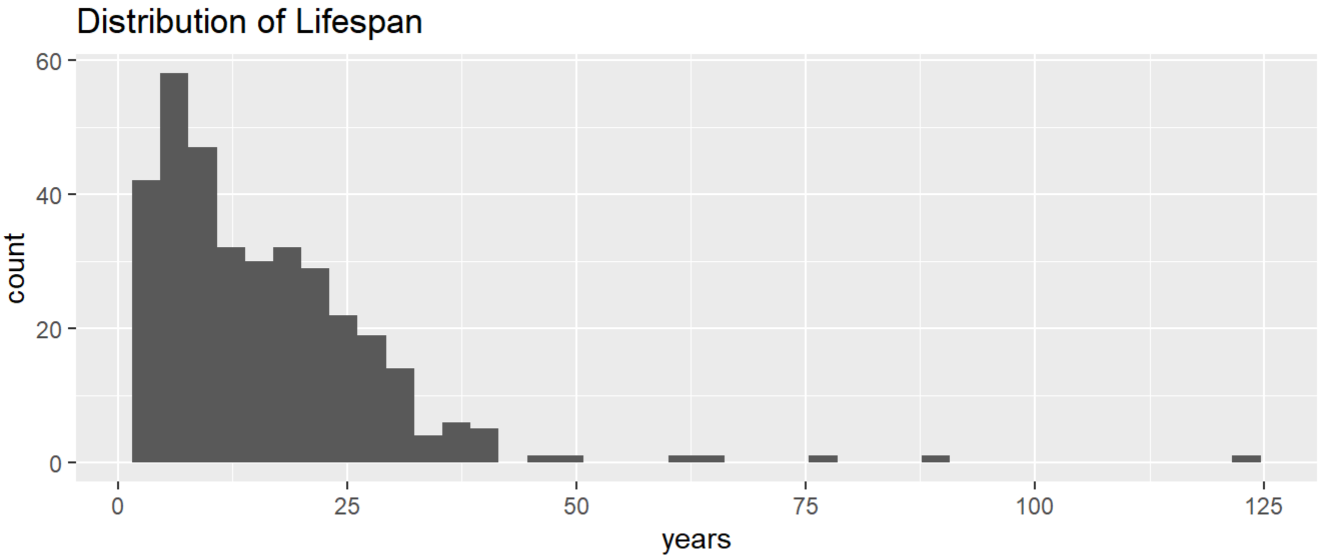

Figure 3: Distribution of animal lifespans

.

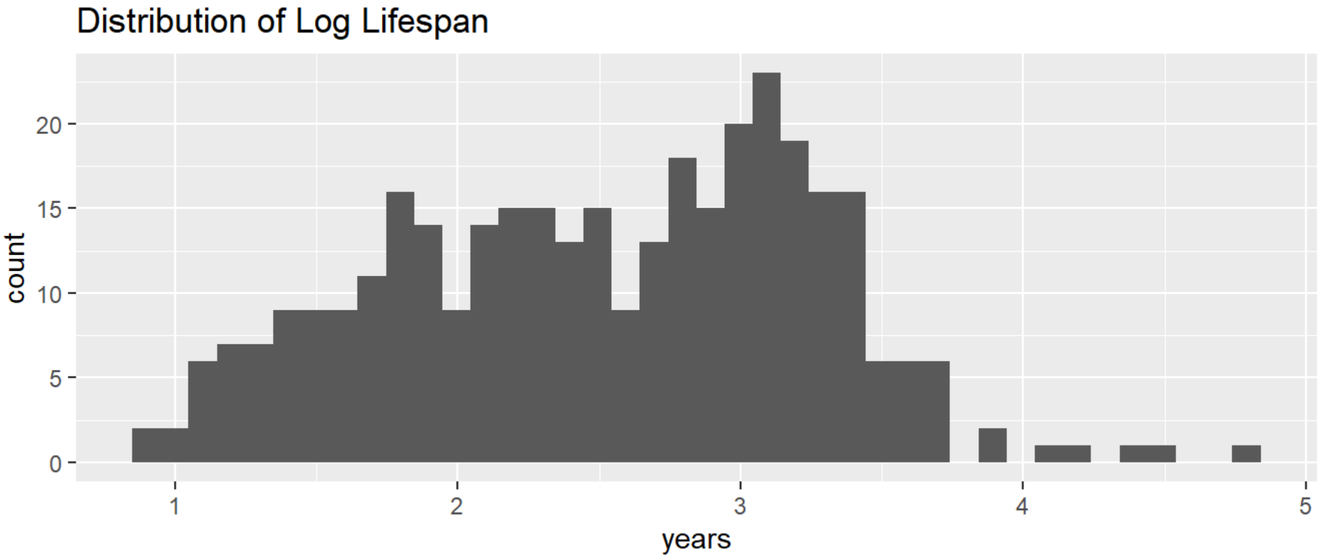

Figure 4: Distribution of log of lifespans

(4) All three tasks only require metabolic rate by mass as the lone explanatory variable. When plotting it against lifespan, we notice a great outlier that represents the data point for the elephant (not shown). This outlier makes it very hard to read the scatterplot in a meaningful way. Rather than removing that point however, it makes more sense to plot the scatter plot of the log transformations of those columns. The result of doing so is pasted in Figure 5. Looking at Figure 5, the points look much more random and follow the best fit line much more closely. This indicates using double log transformation in the model later on may be a good idea.

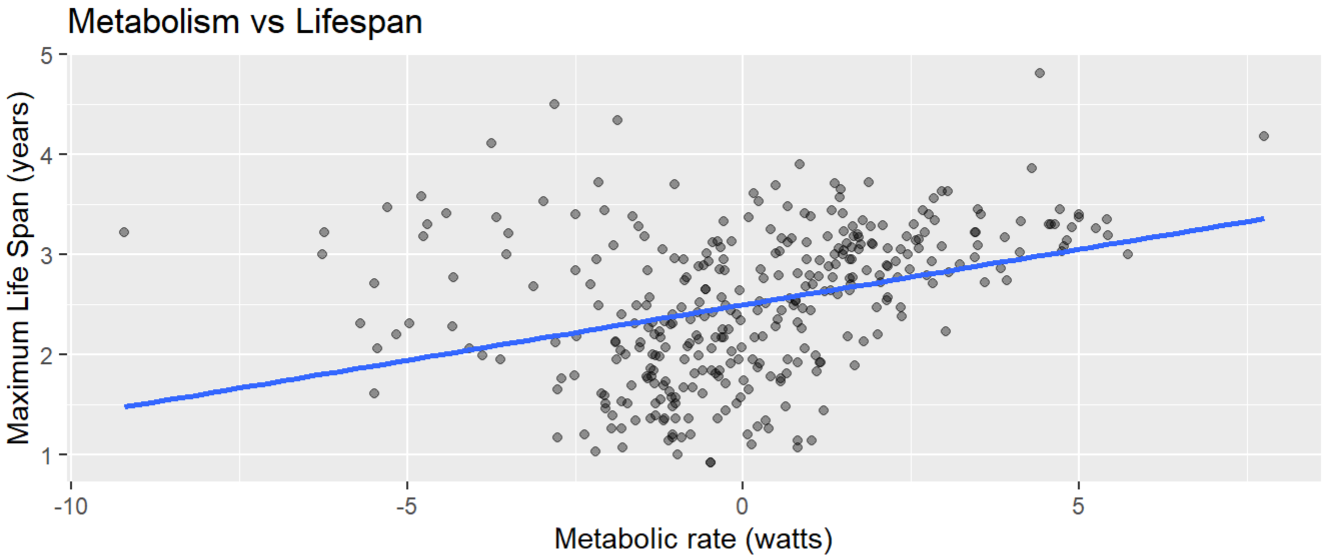

Figure 5: Distribution of log(metabolism) to log(lifespan)

Based on the nature of the problem, it does not seem reasonable to use any of the categorical features, since the classification data is not useful in determining life span. On the other hand, the other quantitative variable may be useful if they have certain trends with metabolism or lifespan. Metabolism and body mass obviously should not be used since there is already a term being used that is based on the operation of those two. **(5)** After producing a correlation of the three leftover measurable variables, we observe that there are no highly correlated variables that results in multi-collinearity.

Based on the correlation coefficient values (not shown), we can assume that metabolism by mass has a weak positive relationships with lifespan, with log transformation improving the positive relationship. **(6)** However, temperature has a weak negative correlation with lifespan. This relationship remains weak after being log-transformed (scatterplot not shown). Because that relationship remains weak, it makes the most sense to not include the body temperature feature in the model moving forward.

Modeling & Diagnostics

(1) The linear model will be made using log transformations to both predictor and response variable. It looks like this:

Linear: $\log(Maximum.longevity.yrs)$ ~ β_0 + $\beta_1 log(Metabolic.by.mass)$ + ϵ

(2) The smoothing spline models using five different df values (3-7) will also be made with the same variables and transformations. The formula looks like the following:

Spline: $\log(Maximum.longevity.yrs)$ ~ β_0 + $s(log(Metabolic.by.mass))$

After the 5-fold cross validation has been done to both the linear and spline models with degree of freedoms 3…7, the errors are outputted in the following table. It seems that the linear model clearly has a higher error than the spline models. The spline models are close in error, even with differences in the degrees of freedom used. **(3)** The pattern between error and degrees of freedom seems to be inverse, with the model having df = 7 returning the smallest error amount. Therefore, it can be assumed to be the best model of the bunch and will be used for future research questions.

	5-fold CV Sample MSE
Linear	178.5

	5-fold CV Sample MSE
Spline df=3	128.5
Spline df=4	122.8
Spline df=5	120.3
Spline df=6	119.4
Spline df=7	119.1

The sample average MSE and SD will also be shown in the Figures 6 and 7 below. In both of these diagrams, the purple values represent the spline values, while the green (only having one point) will represent the linear model. The MSE of the linear model is displayed once for every “df” value so that it can be compared to each respective value for the spline models.

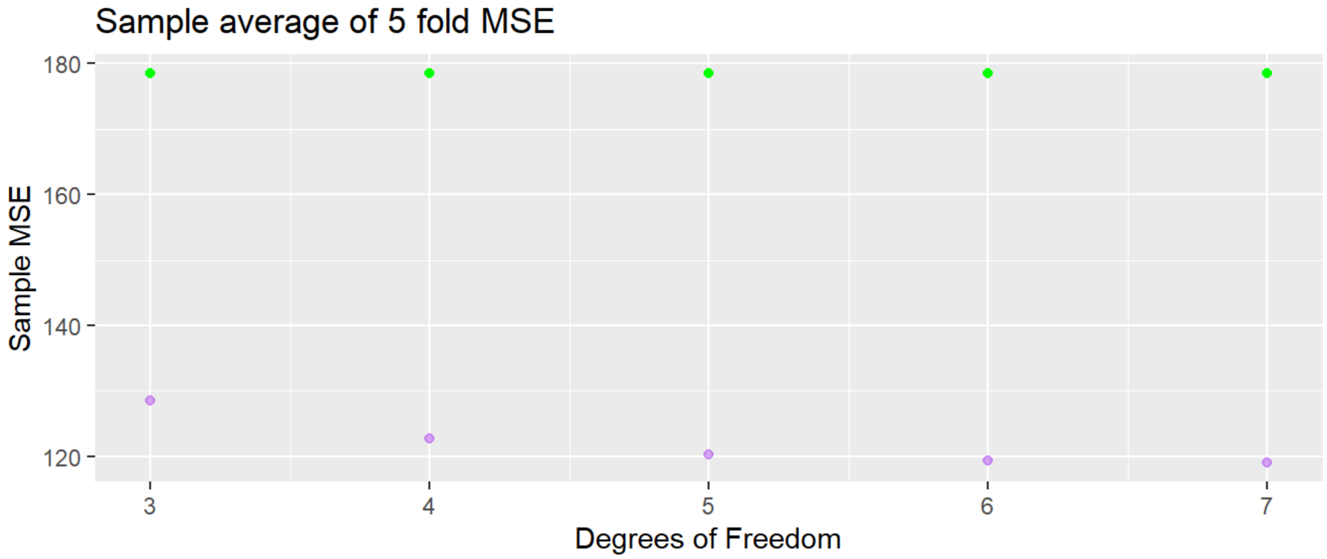

Figure 6: Sample Average of 5-fold MSE for all models

.

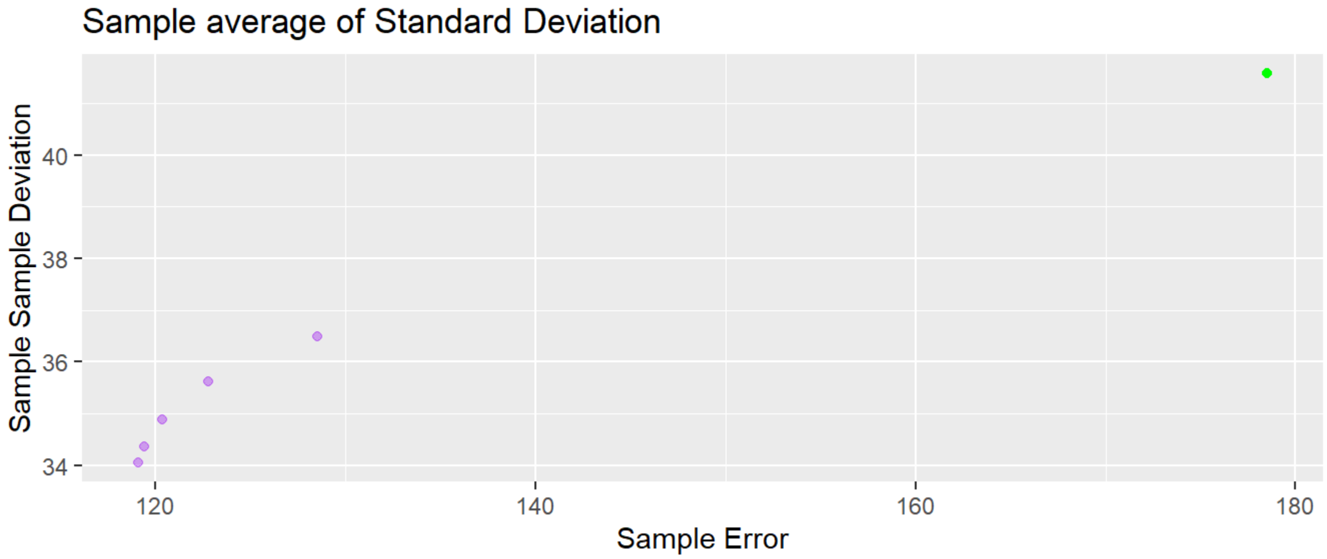

Figure 7: Sample Average of 5-fold Standard Deviation for all models

(4) Based on these three diagnostic models represented by Figure 8 (scatterplot of the residuals), Figure 9 (QQ plot), and Figure 10 (scatterplot compared to best model line), the model assumptions seem mostly satisfied. The first scatterplot shows no patterns and have an average residual of 0. The model pasted onto the next scatterplot looks like a good representation of the data points. The QQ plots looks good as well. with the points following the dotted lines mostly, up until the end points, which may indicate a possible violation in normality. However, it is not significant at all.

Although there is no clear violation of assumptions when looking at both the model and residual diagnostics, it is typically safer to use non-parametric bootstrapping. Since we are only using one predictor variable and both the response and predictor were log-transformed, the predictor may not be as strong as the model indicates. **(6)** Therefore, for the following sections, non-parametric bootstrapping will be used.

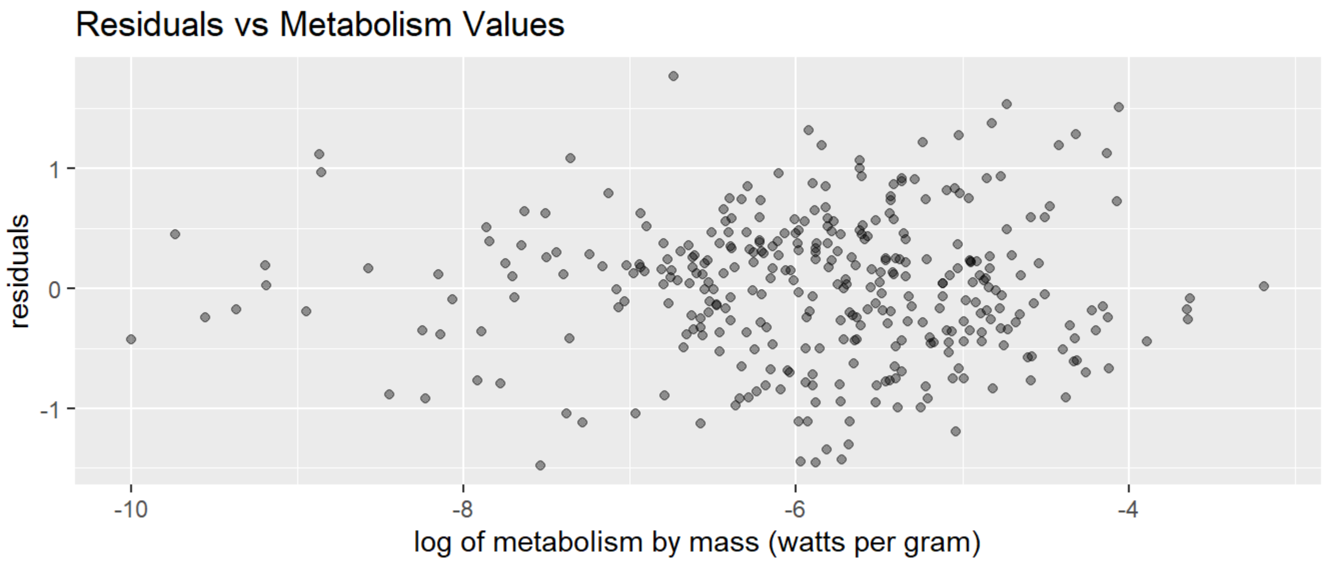

Figure 8: Scatterplot of Residuals for the chosen model (df=7)

.

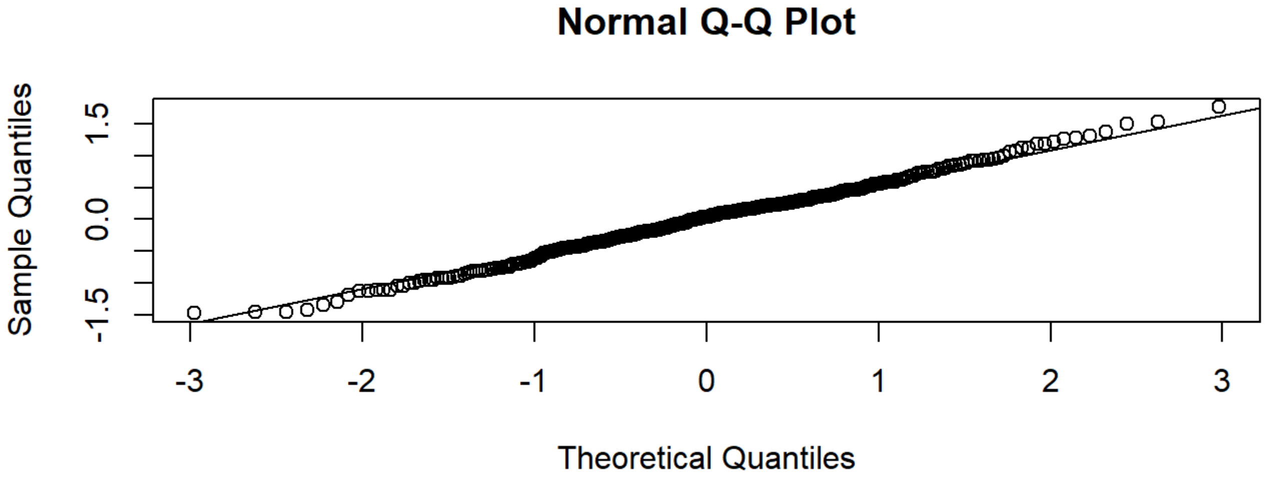

Figure 9: QQ plot for the best model (df=7)

.

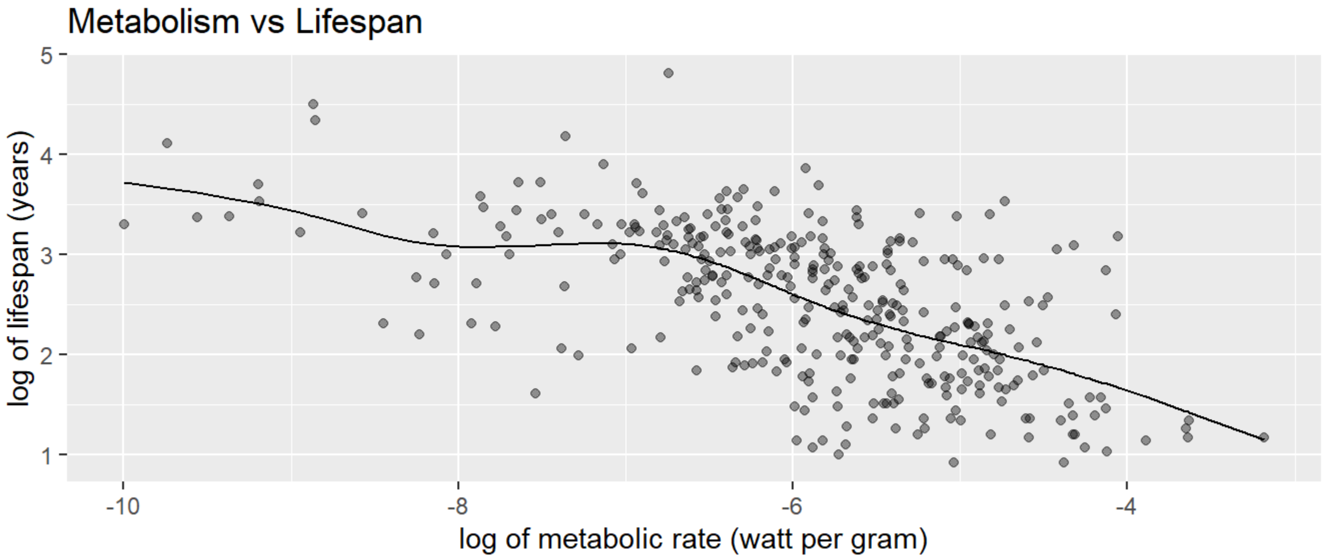

Figure 10: Best model overlayed with EDA scatterplot

(5) Lastly, based on the the uncertainty of the estimates of the prediction error, there does not appear to be a statistically significant different in the spline models when changing the degree of freedom value. However, there does seem to be a difference when comparing the linear model to any of the spline models, as the means square error are much larger.

Results

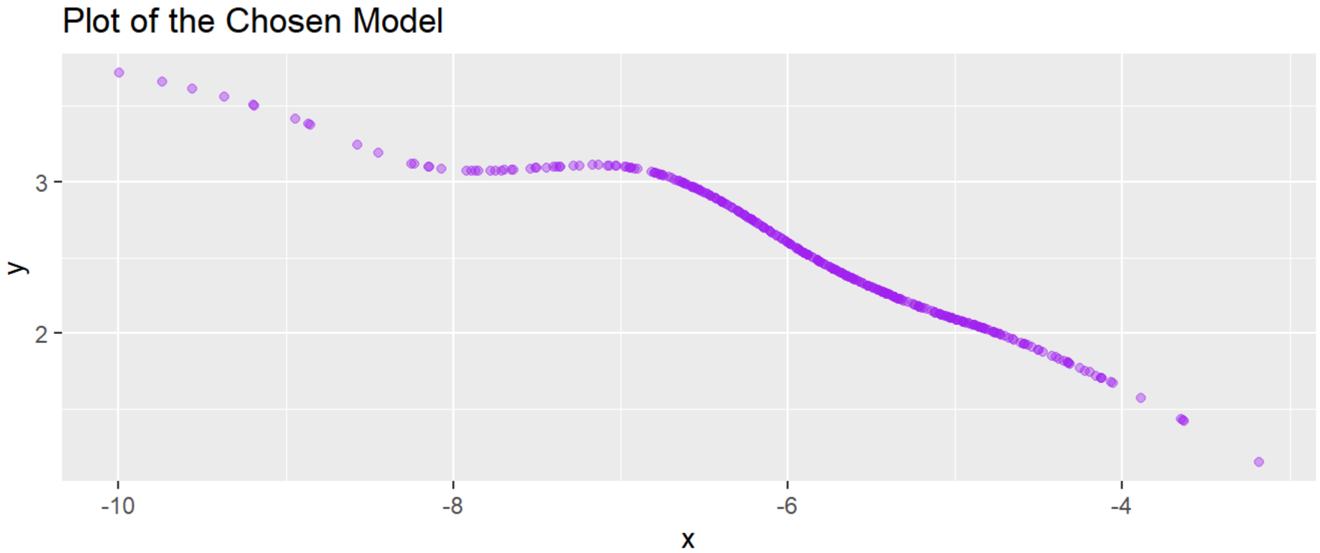

Figure 11: Plot of Spline Model (df = 7)

(1) Based on the trajectory of the chosen model and the scatterplot shown in Figure 10, animals with lower metabolic rates have longer lifespans. This trend is reinforced when plotting the x and y values from the spline model itself, as seen in Figure 11.

	Maximum.longevity.yrs	Body.mass.g	Metabolic.rate	Metabolic.by.mass
83	19	1160	2.588	0.002231

We can apply this finding on to Jorgensen's favorite animal, the crab-eating raccoon. As seen in Table 2, its maximum lifespan is 19 years, while posting a metabolism by mass value of 0.002231 watts per gram. Having a metabolic rate 50% would mean its new 'Metabolic.by.mass' value would be 0.0011155. **(2)** Using the selected model, the lifespan of an animal with a 'Metabolic.by.mass' value is 21.34163.

Overall, when halving the metabolism of the crab-eating raccoon, its predicted lifespan rose 2.34 years, or 12.3%.

However, further analysis can be done. In addition to making single predictions, confidence intervals can also be made. Using a non-parametric bootstrap and a 95% confidence threshold, a confidence interval with 1000 bootstrap iterations can be made for the crab-eating raccoon's estimated lifespan with a halved metabolism. **(3)** After the procedure, the confidence interval can be seen as (19.22811, 24.12923) years.

Since the lower bound is greater than the original value of 19, there is a statistically significant improvement in lifespan using the best chosen spline model with a crab-eating raccoon with half metabolism.

Lower	Upper
19.23	24.13

Conclusions

(1) The main finding between metabolic rate and lifespan is they have a negative correlation. When metabolism decreases, there is evidence to conclude lifespan increases. This correlation is weak without the log transformations, but becomes much stronger when both the predictor and response variable are under a log transformation.

(2) Beff Jezos can indeed conclude that reducing a crab-eating raccoon's metabolic rate by half will cause its lifespan to increase with statistical significance. However, this change is minuscule- (0.22811, 5.12923) years increase over the interval. There is definitely debate as to whether this investment is worth it, even for a billionaire.

Even though the lifespan of the crab-eating raccoon is projected to increase, immortality has not been found. And even if immortality has been discovered, the increase in the raccoon's lifespan is so small that Jorgensen will not be able to enjoy his new-found elixir of life with his favorite animal.

(3) The biggest limitation in the study is the lack of any other predictor features in the models used. Even, when adding 'Metabolic.by.mass', there was only four total numeric features to choose from that may influence lifespan. Furthermore, it did not make any sense to use metabolism or mass as predictor variables since the models already included a feature that is a direct arithmetic operation between the two of them. Lastly, running plots in the EDA section has concluded that 'Temperature' is most likely a weak feature as well, since even the log transformed version of that variable had a weak correlation to lifespan.

It certainly didn't help the temperature values were clustered around the same values- most likely because the vast majority of animals in the data set were mammals. Because the classification of animals in the data set were so skewed, the classification features, which comprised of most of the predictor variables, were functionally useless. Of course, there is no way to log transform categorical data. One hot encoding is possible- but again, the features were really skewed. Not only that, but multi-collinearity was bound to happen since sub-classifications are based on its parent classification.

On the other hand, most of the models worked well and the selected model certainly did not have any violations of assumptions- with the exception of a possible normality violation. However, it wasn't enough to change the outcome of the research.

Lastly, the confidence interval was very close to 19. When rerunning the functions without setting seed, it is possible to get less than 19 as part of the lower bound CI, which completely changes the answer of the research. In that case, the crab-eating raccoon’s lifespan is does not have a statistically significant increase when halving its metabolism.