36-402 DA Exam 1

Yilin Wang (yilinwan)

3/27/2021

Introduction

Is it possible for humankind to achieve immortality? This is a question studied by generations of scientists. While the answer to it remained a mystery after numerous research, a more practical question is often asked: How could we extend the life expectancy for humans? Some recent studies suggest that slowing the metabolic rate could be an answer. According to these studies, a decrease in the metabolic rate might lead to an increase in lifespan. In this report, we will try to explore this idea and answer the following research question: How is lifespan related to metabolic rate? Is the relationship between lifespan and metabolic rate linear? What is the size of the effect that metabolic rate has on the lifespan of animals like crab-eating raccoons? **(1)**?

To answer our research question, we will use data *Anage* from the AnAge Database of Animal Ageing and Longevity, which is a database of animal's aging and life history. *Anage* contains data on lifespan, typical body temperature, typical body mass, and typical metabolic rate for over 4200 species **(2)**.

After extensive analysis, we conclude that animals with slower metabolic rates tend to have a higher life expectancy. According to our analysis, the relationship between metabolic rate and lifespan is not-linear. However, after log-log transformation, we have no evidence to conclude that the relationship between the log of metabolic rate and the log of lifespan is non-linear. To address Mr.Preston Jorgensen's concern, for a crab-eating raccoon whose metabolic rate is 50% lower than its typical value, our model suggests that its maximum lifespan will be 20.78 years, which is 43.2% higher than the expected maximum lifespan of a typical crab-eating raccoon. This assertion merely states an associative fact and does not mean that reducing the metabolic rate of a crab-eating raccoon by 50% will increase its lifespan by 43.2%. However, it is possible to make a more reliable casual claim, that lowering the metabolic rate of a crab-eating raccoon would lead to a 1.9-year increase in its maximum lifespan, under some assumptions. **(3)**

Exploratory Data Analysis

In this report, we will use a part of the *Anage* dataset to answer our research question. We will focus on 5 key variables in the *Anage* dataset. The first one is the response variable $Max.lifespan$. It is defined as the maximum longevity (lifespan) of the animal subjects, measured in years **(3)**. We will use $Max.lifespan$ as a measure of longevity in general. As we are interested in the relationship between metabolic rate and longevity, our major explanatory variable will be *Metabolic.by.mass*, which measures the amount of energy (in Watts) used per gram of body mass. *Metabolic.by.mass* is calculated by dividing the resting metabolic rate of the animal subject by its typical body mass, and it allows us to compare animal subjects of different sizes **(1)**. We have two more supplementary explanatory variables: *Body.mass.g* and *Temperature*. *Body.mass.g* measures the typical body mass of the animal subject in grams, and *Temperature* measures the typical body temperature of the animal subjects in Kelvins. Also, we will include *Class* in our analysis, which allows us to see how the subjects in our dataset are distributed among the 4 classes of animals (Amphibia, Aves, Mammalia, and Reptilia).

We will first examine the key variables individually. Figure 1 shows the distribution of our 4 key variables using histograms. We noticed that the distribution of $Max.lifespan$ is heavily right-skewed, with a mean of 16.01 years and a median of 12.9 years **(3)**. The distribution of *Metabolic.by.mass* also seems to be heavily right-skewed, with a mean of 4.405×10^{-3} Watts per gram and median of 3.023×10^{-3} per gram **(2)**. As for the supplementary explanatory variables, *Temperature* follows a left-skewed distribution with a mean of 308.1 Kelvins and a median of 309.4 Kelvins **(2)**. It is noteworthy that while the *Temperature* for the majority of the animal subjects seems to center around 310 Kelvins, the density of *Temperature* seems to reach a small peak at around 290 Kelvins. This could suggest that we have much fewer cold-blood animal subjects than warm-blood animal subjects in the *Anage* dataset **(6)**. The distribution of *Body.mass.g* seems to be right-skewed, with a few outliers deviates significantly from the center **(2)**. Its mean is 19349 grams and its median is 323 grams, which suggests that it is heavily influenced by outliers. In general, the distributions of all key variables are skewed and suffer from outliers **(2)**.

Having examined the univariate distribution of our key variables, we will now examine how the key variables related to each other. Figure 2 presents a matrix of plots showing the relationship between variables **(4)**. The plots on the diagonal show the marginal distribution curves of the key variables, which is similar to the histograms in figure 1. While the plots on the upper part off the diagonal show the correlation coefficients between

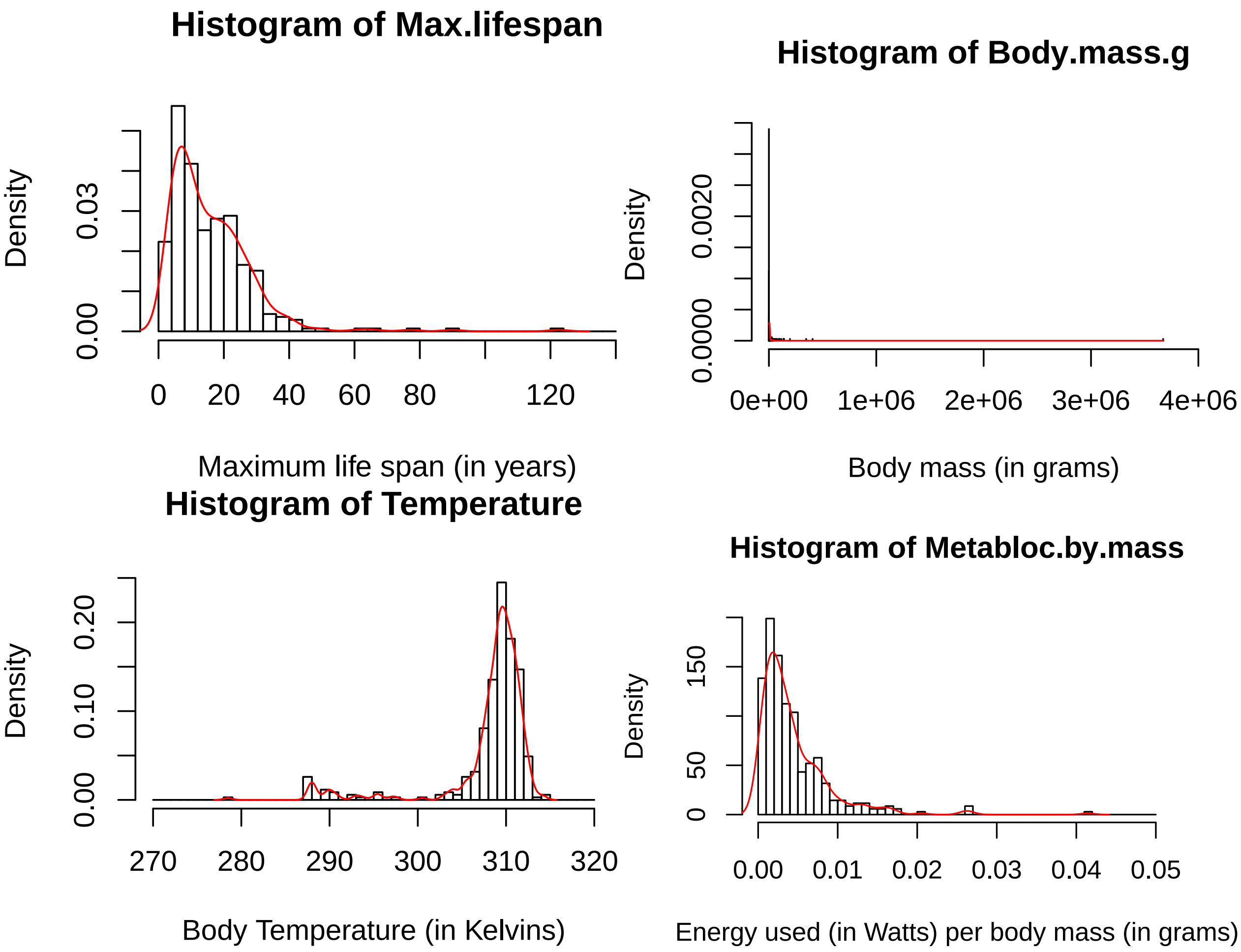


Figure 1: Histogram showing the distribution of Max.lifespan, Body.mass.g, Temperature, Metabolic.by.mass, with estimated density

the key variables, the ones below the diagonal are scatterplots between the key continuous variables. The plots in row 1 are boxplots that show the distribution of *Class* and how the key variables are distributed among the 4 classes **(5)**. We notice that 90.8% of the animals in the Anage data are a member of the class Mammalia (mammals). Although our explanatory variables seem to be related to *Class*, *Class* does not seem to be related to the response variable.

From the scatter plot in figure 2, we see that there seems to be a negative relationship between $Max.lifespan$ and $Metabolic.by.mass$ **(6)**. The negative relationship between the response variable and $Metabolic.by.mass$ is further confirmed by the negative correlation coefficient between them, which is -0.385. However, the relationship between $Max.lifespan$ and $Metabolic.by.mass$ does not seem to be linear, which implies the need for transformation **(6)**. We also see that $Max.lifespan$ seems to be positively associated with $Temperature$

(6). The relationship between these two variables, however, does not look linear as well. From figure 2 and the correlation coefficient (0.027), we also see that *Body.mass.g* seems to be uncorrelated with the response variable *Max.lifespan* **(6)**. It is also noteworthy that *Temperature* seems to be positively associated with Metabolic.by.mass, which suggests that multicollinearity issue may arise in our model **(6)**.

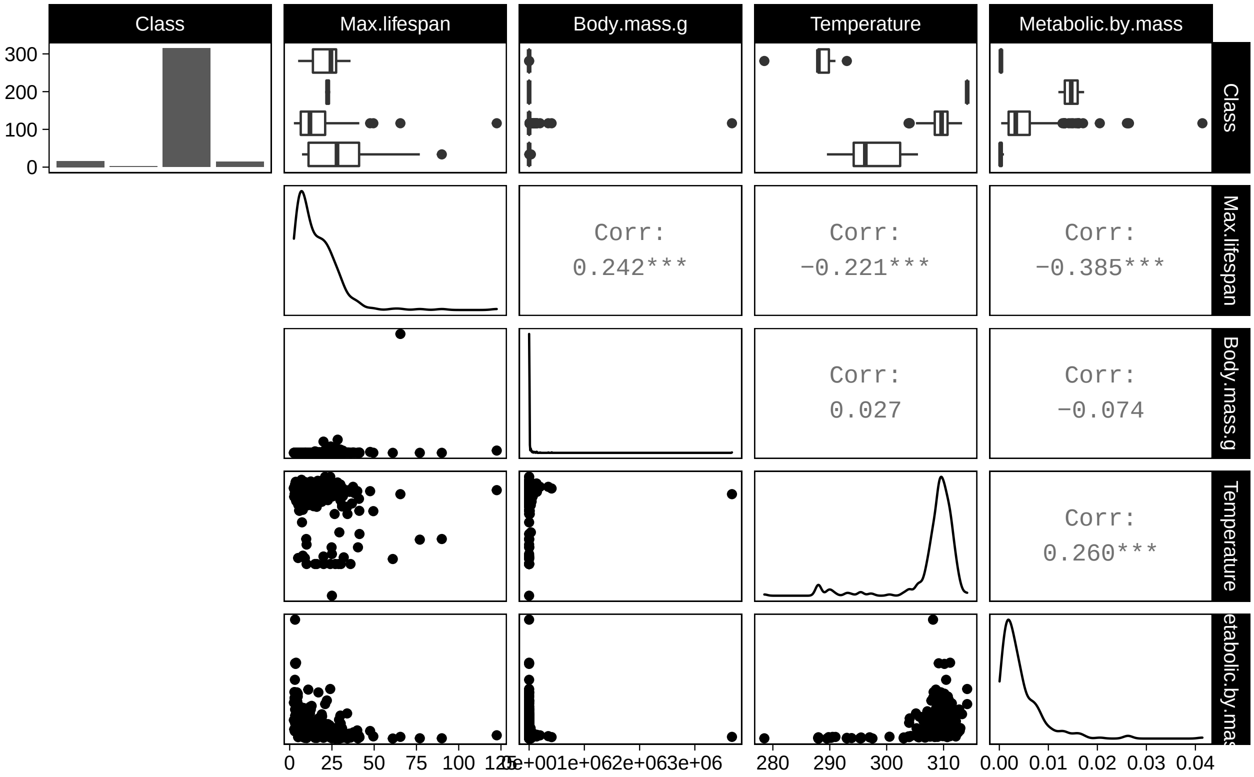


Figure 2: Relationship between key variables, presented in the form of scatter plots, boxplots, and correlation coefficients

We see from the prior analysis that the distributions of the response variable and the major explanatory variable are both right-skewed, and the relationship between them is not linear. This observation suggests that we need to perform transformations on our variables. After several trials, we discovered that applying log transformation on all key variables would be the most appropriate method **(4)**. Customarily, we call this log-log transformation. However, after the transformation, we see in figure 3 that *Temperature* does not seems to be related to *Max.lifespan*. The relationship between them seems weak and is likely due to the effect of outliers **(6)**. From now on, we will be using Age.log = log(Max.lifespan) as the response variable and Metabolic.log = log(Metabolic.by.mass) as the major response variables. From figure 3, we see that the relationship between the *Age.log* and the *Metabolic.log* are linear now **(4)**. Also, *Age.log* and *Metabolic.log* seem to roughly follow the normal distribution, which is a desirable feature **(4)**.

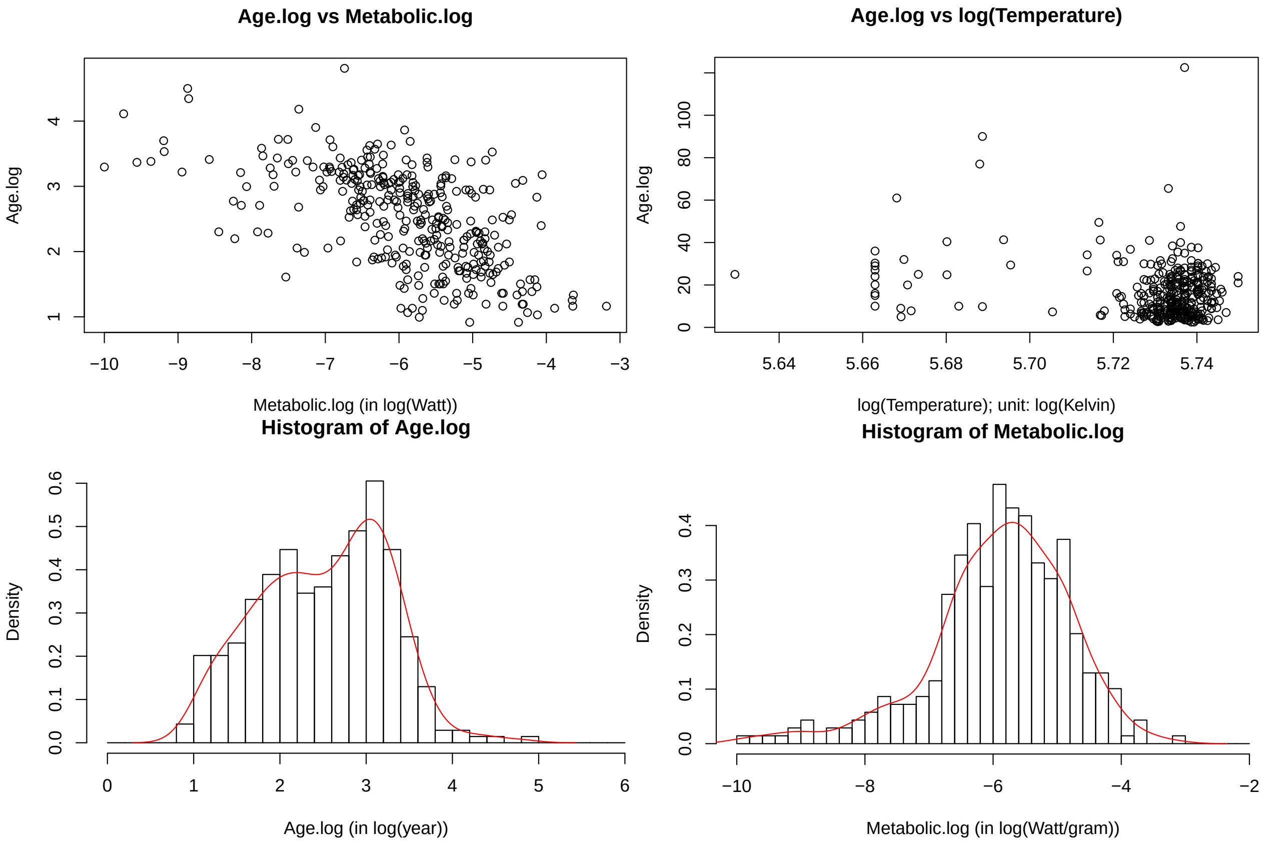


Figure 3: Histograms and Scatter plot between response and explanatory variables after transformation

Modeling & Diagnostics

To answer our research question, we will be building models to explore the relationship between the response variable *Age.log* and the major explanatory variable *Metabolic.log*. More specifically, we will construct our model using linear regression and smooth splines. First, we construct a simple linear regression model, called Model 1, which uses *Metabolic.log* to predict *Age.log*. A summary of Model 1 is as below in table 1 **(1)**:

To make causal inferences, we want to reduce the omitted variable bias by as much as possible. In EDA, we see that all key variables are either related to the response variable or the *Metabolic.log*. Therefore, , we also fit a multiple regression model, which uses all the key variables (*Metabolic.log*, *Body.mass.g*, *Class*, and *Temperature*, all after log transformation) to predict *Age.log*. We call this model Model 1.2, and table 2 present a summary of it. We see that the effect size of *Metabolic.log* is smaller after controlling for

Model:	$Age.log = \beta_0 + \beta_1 \cdot Metabolic.log + \epsilon$
Estimated Model:	$Age.log = 0.14920 - 0.39993 \cdot Metabolic.log$
R-squared	0.342
β_0	0.14920 (SE = 0.17907, p-value = 0.405)
β_1	-0.39993 (SE = 0.02986, p-value $< 2e - 16$)

Table 1: Summary statistics of Model 1

other covariates. Due to the requirement of this report, we will NOT consider Model 1.2 to be a candidate for our final model, but we will use it to gain some insight into the causal relationship between *Metabolic.log* and *Age.log*.

Covariates	$Metabolic.log$, $Class$, $\log(Body.mass.g)$, $\log(Temperature)$
R-squared	0.4552
β_1 **(on Metabolic.log)**	-0.18505 (SE = 0.08981, p-value = 0.0401)

Table 2: Summary statistics of Model 1.2

To allow for more flexibility, we also use splines to construct our model. we will use splines to construct 5 models to predict *Age.log* by *Metabolic.log*. For each model, we will use degrees of freedom (df) = 3, 4, 5, 6, 7, respectively, and we will refer to them as Model 2, Model 3, $\cdots$, Model 6 (Model 2 has $df = 3$ and Model 6 has $df = 7$) **(2)**.

We often prefer simpler models as more flexible models will often over-fit the data and have poor out-of-sample performance. Therefore, we will use cross-validation (CV) to perform model selection. More specifically, we will compute the leave-one-out cross-validation (LOOCV) error for each model, and we will pick the model that minimizes the LOOCV error. Table 3 presents the LOOCV error for Model 1 to Model 6 respectively **(3)**. In table 3, we also provide the standard deviation (SD) and the standard error (SE) of the LOOCV error. We notice that Model 1.2 minimizes the LOOCV error. However, as we do not consider it a candidate for the final model, we will choose Model 5 (spline with $df = 6$), which gives the second least LOOCV error, as our final model. **(3)**

Although Model 5 has the least LOOCV error among the 6 models, its performance is not significantly better than the rest 5 models **(5)**. As the difference in LOOCV error (which is around 0.001 to 0.01) between Model 5 and the rest 5 models are much smaller than the standard error (SE) of the models (which is around 0.027), the difference between model 5 (or Model 1.2) and the rest 5 models is not significant in terms of LOOCV error. Thus, we

Model	Type	df	LOOCV Error	SD(error)	SE(error)
Model 1	SLR	2	0.3665708	0.4882068	0.02620831
Model 1.2 **(*)**	MLR	7	0.3102558	0.4613869	0.02476855
Model 2	Spline	3	0.3581313	0.4999202	0.02683712
Model 3	Spline	4	0.3561909	0.5042282	0.02706839
Model 4	Spline	5	0.3551012	0.5021727	0.02695804
Model 5	Spline	6	0.3546276	0.4988142	0.02677775
Model 6	Spline	7	0.3548524	0.4962062	0.02663774

Table 3: Leave-one-out cross validation (LOOCV) error for model 1 - 6, with standard deviation (SD) and standard error (SE). $SE = SD/\sqrt{n}$ and n is the number of observations in the data. SLR stands for Simple Linear Regression and MLR stands for Multiple Linear Regression. **(*)**: Model 1.2 is excluded from the model selection process (see *Appendix*)

have no evidence to conclude that the relationship between *Metabolic.log* and *Age.log* is not linear. Further, if interpretability and simplicity are prioritized, we could pick Model 1 instead of Model 5 as our final model.

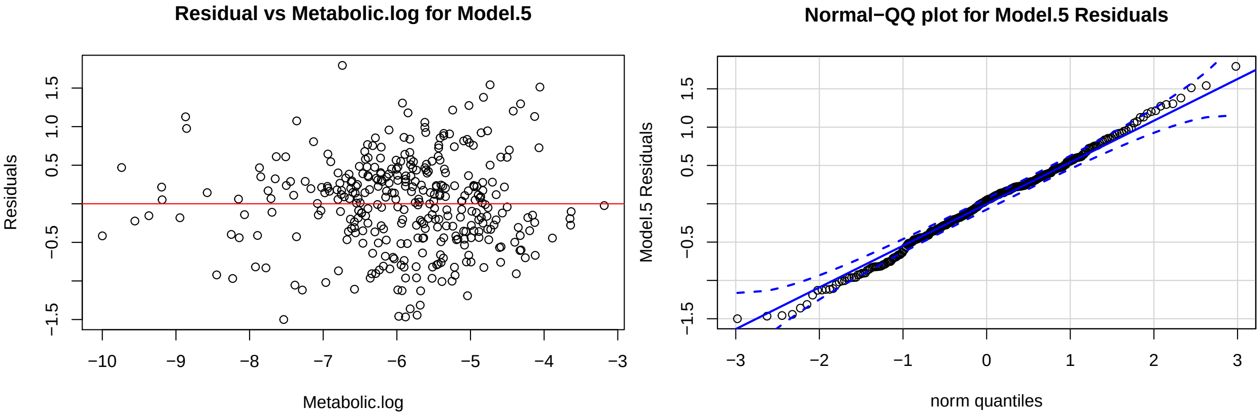


Figure 4: Residual lot and normal-QQ plot of Model.5

Having selected the model, we also want to check whether the assumptions of our model have been violated. Figure 4 presents the residual plot (residual vs Metabolic.log) and the normal-QQ plot for Model 5. From the residual plot, we could conclude that our model fits well and successfully captures the patterns in the data, as the residuals appear to be patternless **(4)**. The residuals are also closely distributed along the theoretical normal quantiles, which suggests that the residuals roughly follow Gaussian distribution.

However, homoscedasticity seems to be slightly violated as residuals seem to have different spread **(4)**. The variance of the residuals seems to be larger when Metabolic.log is greater. To address this issue, we may consider applying better transformations to our variables so that homoscedasticity would hold **(4)**. A more practical method, on the other hand, would be to use weighted least squares (WLS) instead of ordinal least squares (OLS) in model construction **(4)**. More specifically, we give less weight to observations with larger variance as they provide less information. It would also be helpful to use robust variance measures (like sandwich variance, BMB) in the inference **(4)**. For this model, as we have noticed that the variance of residuals depends on the explanatory variable Metabolic.log, re-sample cases (non-parametric bootstrap) would be the most appropriate method to use when we want to calculate relevant statistic's **(6)**. Though the violation of homoscedasticity is not severe, we use non-parametric bootstrap for cautious concerns.

Results

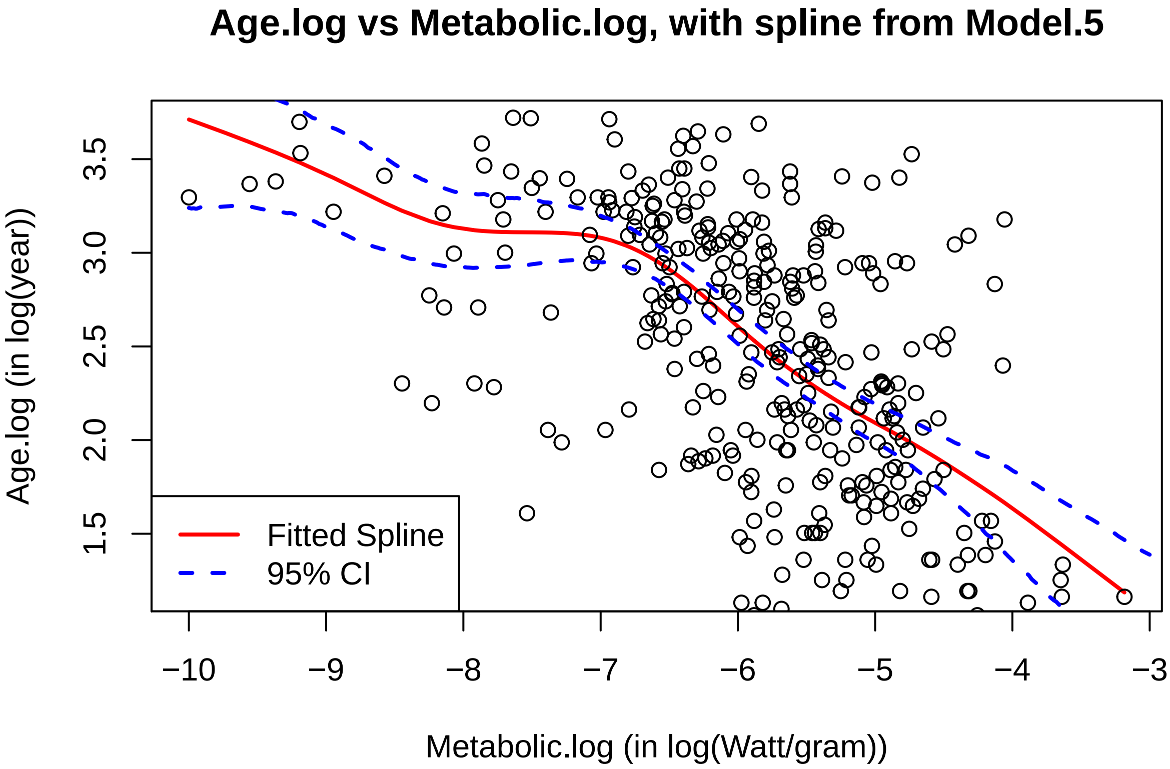


Figure 5: Age.log vs Metabolic.log, with fitted spline from Model.5 and 95 percent confidence interval estimated using nonparametric bootstrap

Having constructed the model, we can now answer our research question. Figure 5 visualizes our model by overlaying the fitted spline (from Model 5) on top of the scatter plot between *Age.log* and *Metabolic.log*. A 95% confidence interval for the prediction is also shown on the graph. We notice that the red line, which represented the fitted spline from Model 5, is downward-sloping. We know that log transformation is a monotone transformation, and it preserves the order of the input. In other words, $\log(a) > \log(b)$ would imply that $a > b$. Therefore, we conclude that metabolic rate is negatively associated with the maximum lifespan of the animal subjects in our data **(1)**. Also, we conclude that the relationship between metabolic rate and maximum lifespan is non-linear.

We could also make a stronger casual claim using Model 1.2, under the assumption that *Class*, *Body.mass.g*, and *Temperature* are the only three confounding variables in the relationship between *Age.log* and *Metabolic.log*. Under such assumptions, we claim that lower metabolic rates lead to higher maximum lifespans **(1)**. The coefficient on *Metabolic.log*, which is -0.18505, is negative. We want to test whether this coefficient is less than 0. As no assumptions are severely violated for Model 1.2 (see *Appendix*), we could directly use the p-value (which is 0.0401) returned by the regression output. As the p-value for the coefficient on *Metabolic.log* is less than 0.05, we conclude that lower metabolic rates lead to higher maximum lifespans, under prior assumptions. Under the prior assumption, we could also say that holding *Class*, *Body.mass.g*, and *Temperature* constant, decrease metabolic rate by 1% increase maximum lifespan by 0.1851%.

We will use data on crab-eating raccoons to illustrate our findings. Crab-eating raccoon is an animal whose metabolic rate is 2.588 Watts and whose typical body mass is 1160 grams **(2)**. The maximum lifespan of the crab-eating raccoon is 19 years in our dataset **(2)**. For a typical crab-eating raccoon, Model 5 predicts that its maximum lifespan is expected to be 14.51303 years. On the other hand, for a crab-eating raccoon whose metabolic rate is 50% less than its typical value, our model predicts that its expected maximum lifespan is expected to be 20.77763 years, which is 43.2% higher than that of a typical crab-eating raccoon and 9.35% higher than that of the crab-eating raccoons in our data **(2)**.

We computed a 95 % pivotal confidence interval (CI) for the prior prediction via bootstrapping. For each of the 1000 rounds of bootstrap, we first construct the model under the bootstrapped sample, and we estimate Age.log of the crab-eating raccoon whose metabolic rate is 50% less than its typical value. We then take the exponential of that estimate as our prediction of the maximum lifespan for that round of bootstrap. Table 4 summarizes the pivotal CI's for the prediction of Age.log and Max.lifespan respectively, using Model 1.2 and Model 5.

Model 5 does not control for any potential confounding variables, which means that casual claims drawn based on Model 5 are likely to be invalid. Compared to Model 5, Model

Model	**Variable**	**Prediction**	**95% Confidence interval**
Model 5	Age.log	3.033877	[2.920867, 3.137564]
Model 5	Max.lifespan	20.77763	[18.29174, 22.82408]
Model 1.2	Age.log	2.757543	[2.590037, 2.916673]
Model 1.2	Max.lifespan	15.76107	[12.88700, 18.07975]

Table 4: Summary of pivotal confidence intervals

1.2 controls for more potential confounding variables and are more reliable for casual inference. Using Model 1.2, we predict that the maximum life expectancy of a typical crab eating raccoon is 13.86 years, and the maximum life expectancy is 15.76 years for a crab-eating raccoon whose metabolic rate is 50% less than its typical value. If we assume that $Class$, $Body.mass.g$, and $Temperature$ are the only three confounding variables, then we could claim that reducing the metabolic rate of a crab-eating raccoon will increase its maximum lifespan by 1.9 years.

Conclusions

From prior analysis, we conclude that animals with lower metabolic rates have higher lifespans, on average **(1)**. While this is only an associative statement that does not involve any causal claims, we can conclude that lower metabolic rate leads to a higher maximum lifespan under the assumption that class, typical body mass, and typical body temperature are the only confounding variables in the relationship between metabolic rate and maximum lifespan **(1)**. The relationship between metabolic rate and maximum lifespan is not linear **(1)**. However, after log-log transformation, we have no evidence to conclude that the relationship between $Age.log$ and $Metabolic.log$ is not linear **(1)**. Though our final model, which minimizes LOOCV error, uses a smooth spline with $df = 6$, its LOOCV error is not significantly lower than that of the simple linear model.

As for Mr.Jorgensen, our model suggests that a crab-eating raccoon whose metabolic rate is 50% less than its typical value is expected to have a maximum lifespan of 20.77763 years, which is 43.2% higher than that for a typical crab-eating raccoon **(2)**. However, there is no guarantee that reducing the metabolic rate of a crab-eating raccoon by 50% will cause its lifespan to increase by 43.2% **(2)**. As our final model does not control for all confounding variables, casual claims derived from it are likely to be invalid. A more reliable casual statement, on the other hand, could be drawn using Model 1.2. Under the assumption that there are no more confounding variables except for *Class*, *Body.mass.g*, and *Temperature*, we can claim that reducing the metabolic rate of a crab-eating raccoon would lead to 1.9 years increase in its maximum life expectancy **(2)**.

Despite Model 1.2 mitigates some degree of omitted variable bias, unobserved confounding variables may still exist, which means that the true causal relationship could differ from our model estimate**(3)**. For example, herbivore animals may tend to have lower metabolic rates and longer lifespans than carnivore animals, which resulted in the negative relationship between lifespan and the metabolic rate we see in our model. Under this premise, changing the metabolic rate of a crab-eating raccoon would not affect its lifespan as it does not change its dietary habit. Also, in this model, we use maximum lifespan, instead of average lifespan, as a measure of life expectancy. Maximum lifespan is more easily influenced by outliers and is less representative of the life expectancy of the population **(3)**. Also, maximum lifespan is an aggregate measure of life expectancy, which means that we only have 1 observation for most species. Therefore, our model could fail to capture some crucial patterns in the data (e.g., how the metabolic rate is related to lifespan within a certain species) **(3)**. For further research, the study should try to include more potential confounding variables in the model. Also, it should use a dataset that measures the lifespan of every individual subject.

Appendix and Notes

1. We don't consider Model 1.2 as a candidate for the model selection process because the rubric explicitly asks us to use binary regression. However, we include it in the analysis as we feel that it's essential for us to answer the casual question.

2. As we transformed the response variable, we predict $Age.log$ first, and we then take the exponential of that as the estimate of $Max.lifespan$.

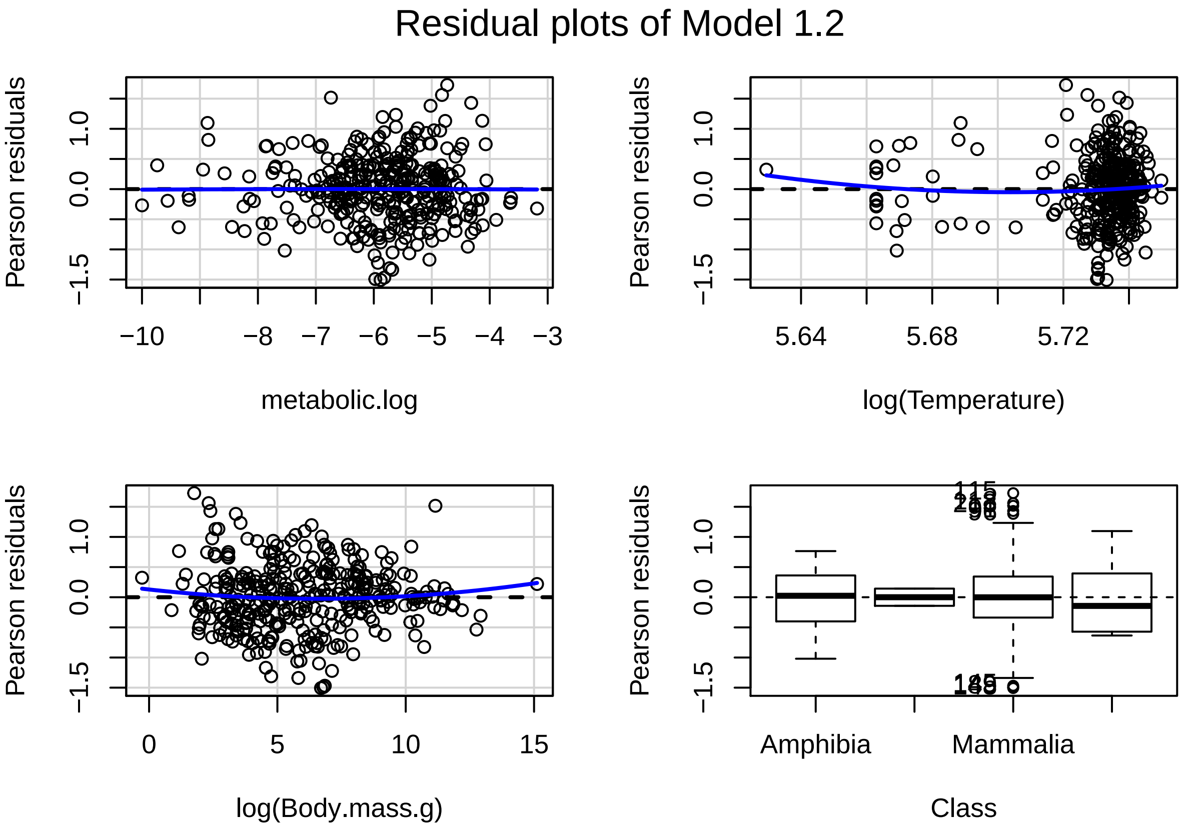
3. The diagnostic plots of Model 1.2 are on the next page. We see that the residuals seem to be patternless and have the same spread, which suggests that linearity and homoscedasticity hold. Also, the residuals are closely distributed along the normal quantile, which suggests that the Gaussian error assumption holds. Thus, no assumptions are severely violated for model 1.2, which means we could directly use the p-value returned by the regression output.

4. As the homoscedasticity is only slightly violated and the Gaussian error assumption holds for Model 5, we could also use parametric bootstrap to calculate the confidence

interval. As no assumptions are violated for Model 1.2, we could (should) also use parametric bootstrap. We used non-parametric bootstrap to calculate the CI for Model 1.2 to be consistent with Model 5. Table 5 (the following table) are the CI's calculated using parametric bootstrap. We notice that the CI's are generally narrower.

Model	Variable	Prediction	95% Confidence interval
Model 5	Age.log	3.033877	[2.953717, 3.191427]
Model 5	Max.lifespan	20.77763	[19.04351, 23.80631]
Model 1.2	Age.log	2.757543	[2.58517, 2.920729]
Model 1.2	Max.lifespan	15.76107	[12.79608, 18.13417]

Table 5: Summary of pivotal confidence intervals (Parametric Bootstrap)



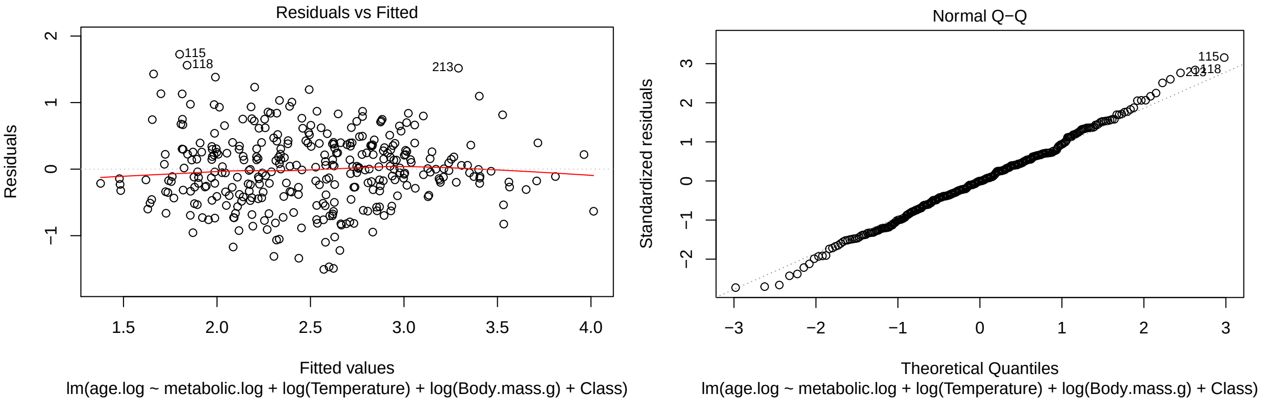


Figure 6: Diagnostic plots for Model 1.2