36-402 DA Exam 1

Jacky Liu (jackyl1)

3/19/2021

INSTRUCTIONS – REMOVE BEFORE SUBMITTING

This is a template for your data analysis report. It should work for anyone who is using R Markdown to generate PDFs through LaTeX. If you make the PDFs in some other way, as long as you can get 12 point fonts and reasonable margins, it will be fine.

Marking your answers

Each section below contains numbered questions. You are **required** to mark the sentence in your report that most closely answers each question.

For example, for question 2 in the EDA section, mark the sentence with **(2)**. For question 3 in Modeling & Diagnostics, mark the sentence with **(3)**.

You are writing a report, not simply writing answers to the questions, so you should **not** leave the questions in your report. Nor should your report consist only of bullet points or a numbered list of answers.

Figure captioning/sizing tips

Here is an example of creating a plot, setting its size (in inches), and giving it a caption. Figures with captions "float" on the page, usually appearing at the top or bottom; don't try to fight LaTeX and force them to appear in a certain place, because you will lose.

Notice that the figure appears in the PDF, but the code does not.

You can only make one figure per code chunk.

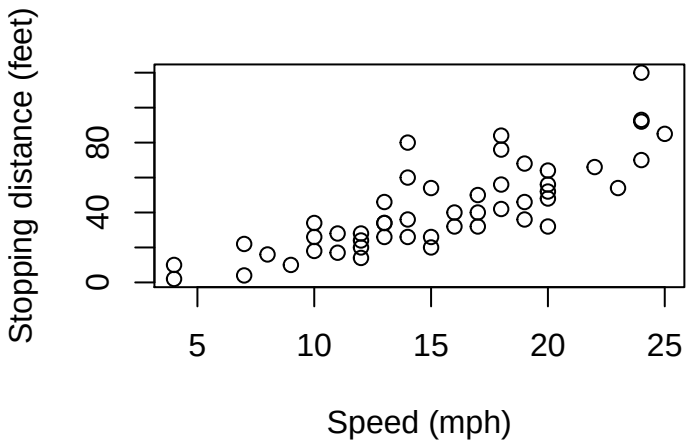

Figure 1: Distance it took cars at each speed to stop.

Introduction

1. Clearly state the research questions and objectives of your study.
2. State what data you are using and what information it contains.
3. Briefly mention your final findings. State them in terms that Preston Jorgensen can understand, rather than using technical jargon.

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.

Exploratory Data Analysis

1. Create a new variable that divides `Metabolic.rate` by `Body.mass.g`. Call this `Metabolic.by.mass`. This is the amount of energy used per unit of body mass, and will allow comparisons between animals of different sizes. Use this variable instead of `Metabolic.rate` in the rest of your report.

2. Explore the key variables you need to answer the research questions. Describe them with any necessary univariate EDA, such as plots or summary statistics.

3. Identify and specify your response variable and its distribution.

4. Do multivariate EDA to explore the relationship between predictors and the response. Based on this and your univariate EDA, are there transformations you should use before modeling your data? After any transformations, do the relationships appear to be linear? If you decide to use transforma tions, use them for the rest of your analysis.

5. If there are plots that would help answer the research questions, by showing key relationships, show those plots.

6. Describe any trends or interesting features you see that suggest what you will find in the 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. 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.

From the figure above, (3) we can see that our response variable, which we defined as lifespan, has a very right-skew distribution. The majority of animals we are looking at

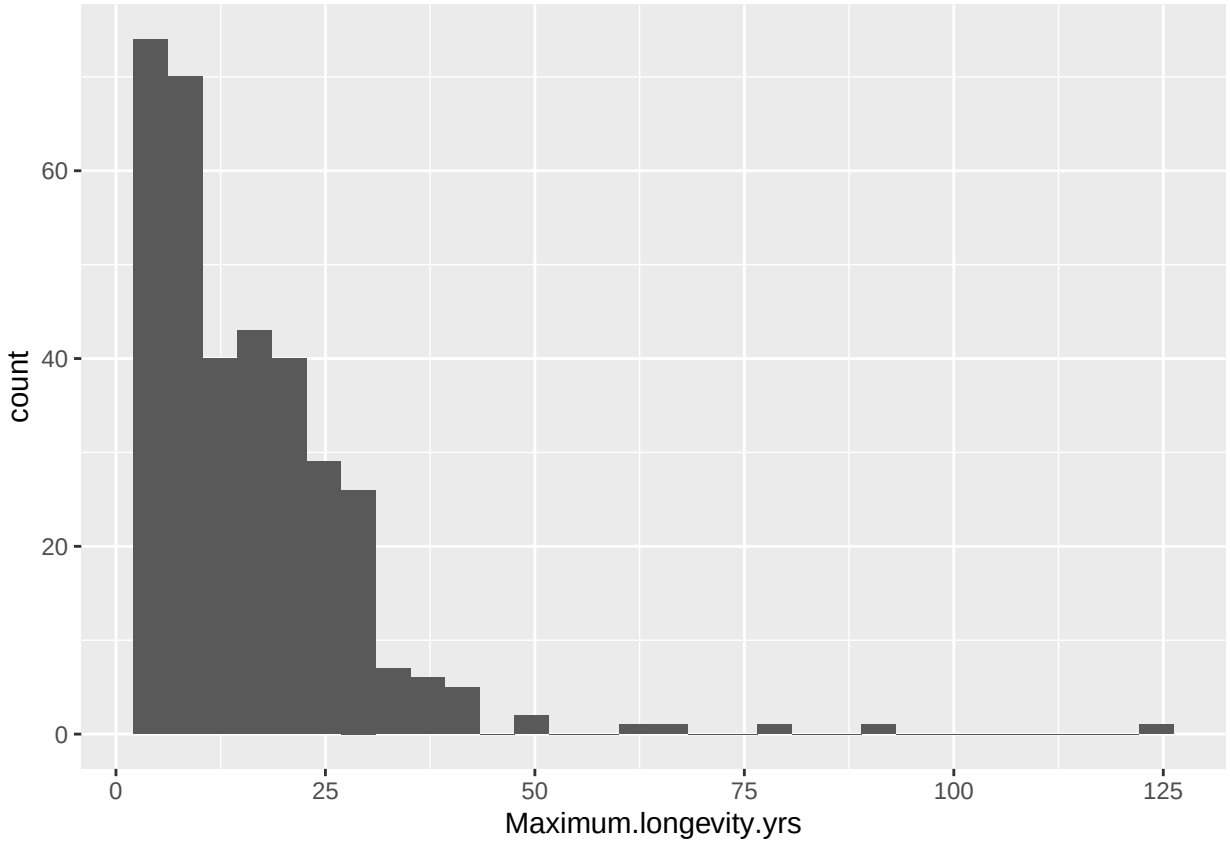

Figure 2: Histogram showing marginal distribution of our response variable, lifespan.

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. 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.

Modeling & Diagnostics

1. Construct a linear model to predict `Maximum.longevity.yrs` using `Metabolic.by.mass`. (If you decided to use transformations of either variable, use those when fitting your model.)
2. Because you're working for a billionaire, you'd like to show off a more sophisticated and flexible model. Use a smoothing spline to fit to the same data. Fit five models, setting the `df` (effective degrees of freedom) to be 3, 4, . . . , 7.
3. Use cross-validation to determine which of the models fit best to the data, in terms of prediction error. Your comparison should include the linear model and the spline models. (You can use any type of cross-validation you think is suitable.) State the error you estimated for each model and state which model you decided is best.
4. Present model diagnostic plots for your selected model. Discuss whether the model appears to fit well, and describe any possible improvements to your model to address any violations of the model assumptions.
5. Comment on whether the difference between the models appears significant, based on the uncertainty in your estimates of the prediction error; a formal test is not required here.
6. For your chosen model, examine the residual diagnostics to determine what type of bootstrap would be appropriate for this data.

Results

1. Using the selected model, determine whether animals with lower metabolic rates have longer lifespans, as requested by Preston Jorgensen. You can use model coeffi-

cients, tests, or plots to answer this question.
2. Look up the crab-eating raccoon in the data and obtain its features. Calculate what metabolic rate it would have it its rate were 50% smaller. Use your best model to estimate the mean lifespan of an animal with those characteristics.
3. Use bootstrapping to make a 95% confidence interval for that quantity. Use the pivotal confidence interval and 1000 bootstrap iterations. Report the confidence interval.

Conclusions

1. Summarize your main findings about the relationship between metabolic rate and lifespan.
2. Explain whether Preston Jorgensen can conclude that reducing the crab-eating raccoon's metabolic rate by 50% would cause its lifespan to change, and if so, by how much.
3. Discuss any limitations to your analysis that might affect the conclusions, such as violations of assumptions or limitations in the data.