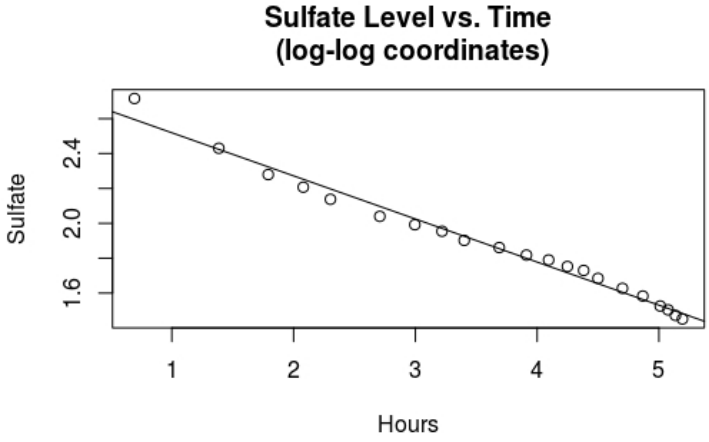
REPORT

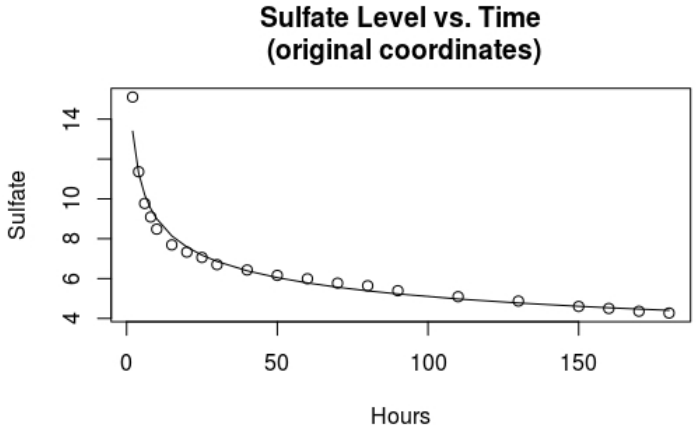
I have written code based on, what appears to me, all the requirements, and it runs without errors on my machine. Below are the results with my explanation.

Part 1

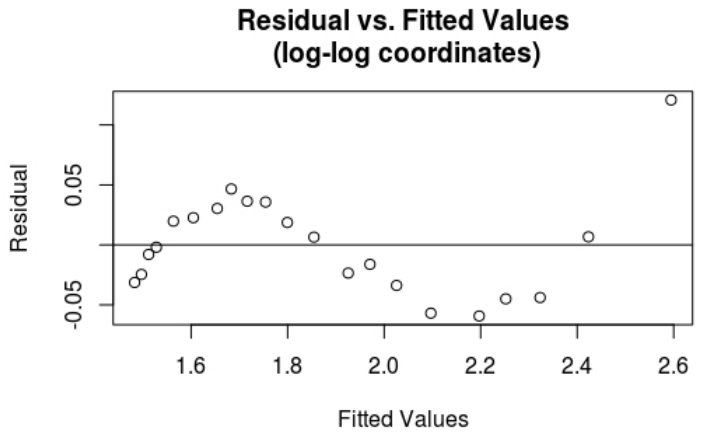
a) Plot showing (a) the data points and (b) the regression line in log-log coordinates:

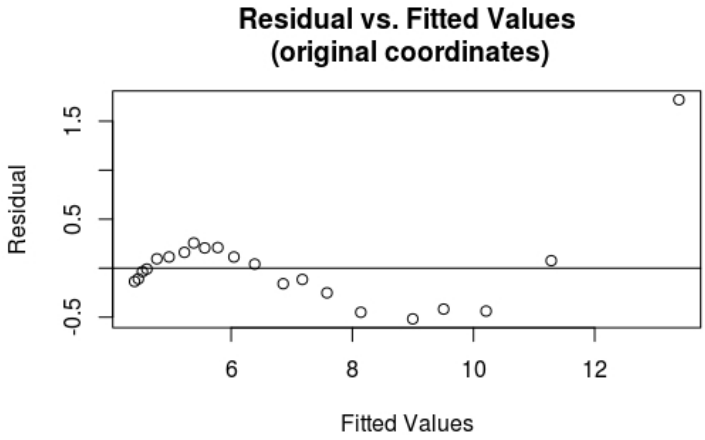


b) Plot showing (a) the data points and (b) the regression curve in the original coordinates:



c) Plot the residual against the fitted values in log-log and in original coordinates:

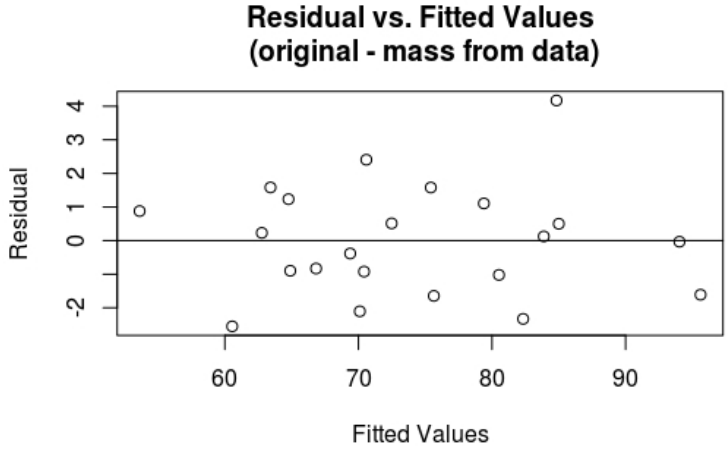




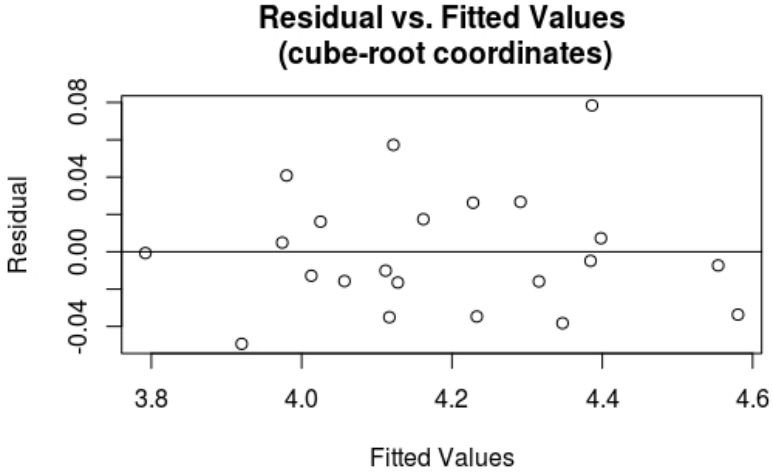
d) The regression line fits the data more or less well, but the plot showing Residuals vs. Fitted Values has an interesting pattern on it, almost like a sinusoidal function. This can be a sign of the fact that our regression is not exactly linear.
As suggested here https://onlinecourses.science.psu.edu/stat501/node/279, there is a relationship between the sulfate content and time, but it is not linear because the residuals depart from 0 in a systematic manner forming a sinusoid. So, this relationship would have been better described by a non-linear model. The r2 value for this model is very high (0.983). According to the same source, this should not be interpreted as the regression line fitting the data well. This just means that the sulfate levels are better predicted when time is considered rather than when it is not considered. But the last two plots say that the prediction would have been better if I had used a non-linear model.

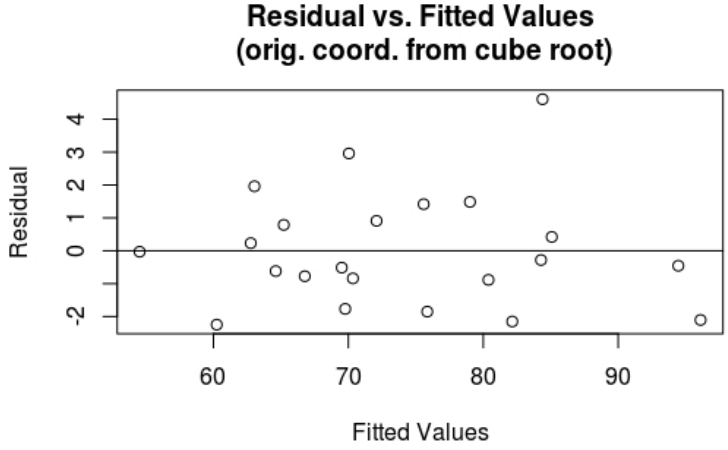
Part 2

a) Linear regression of predicting the body mass from diameters. Plot of residual vs. fitted values:



b) Linear regression of cube root of mass against diameters. Plot of residual vs. fitted values in cube-root coordinates and in original coordinates:



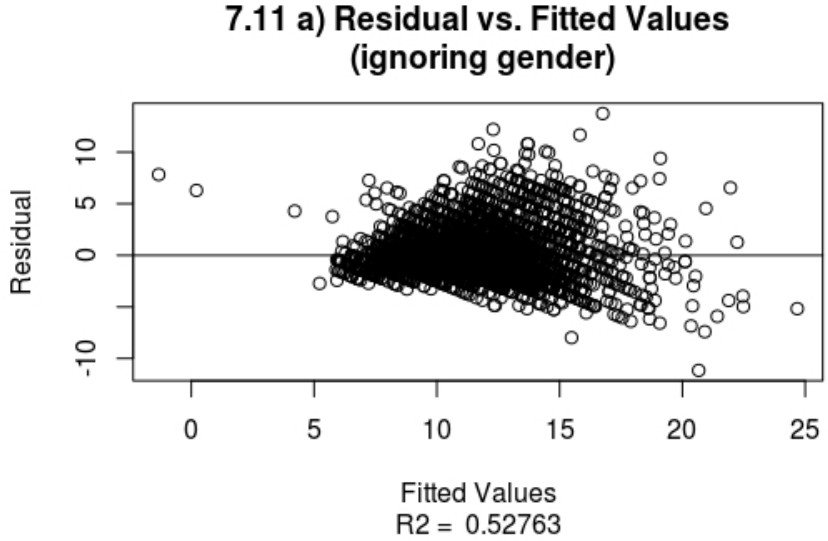


c) In b) we're doing a linear regression in a non-linear (cube-root) space which means we are learning a different function than in (a). Thus, the residuals should look different too. This fact is confirmed by the 1st and 3rd plots above (original coordinates). They seem to be very similar, but if you look close, especially if you toggle between the two figures in a graphic viewer (graphic files are attached to this submission), you can see differences in the left portion of the graph, some in the center and in the right portion. However, these differences are not significant which is also confirmed by a statement on the dataset's website (although for a different type of transformation) saying that one can transform all the variables, but "the original scales are very nearly as good."

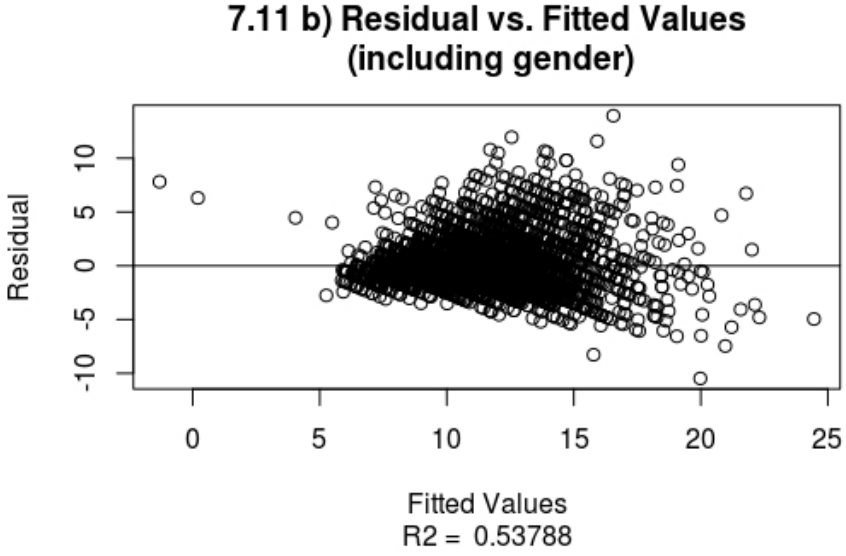
To answer the question which regression is better – they all look similar, so they all should be more or less equally good.

Part 3

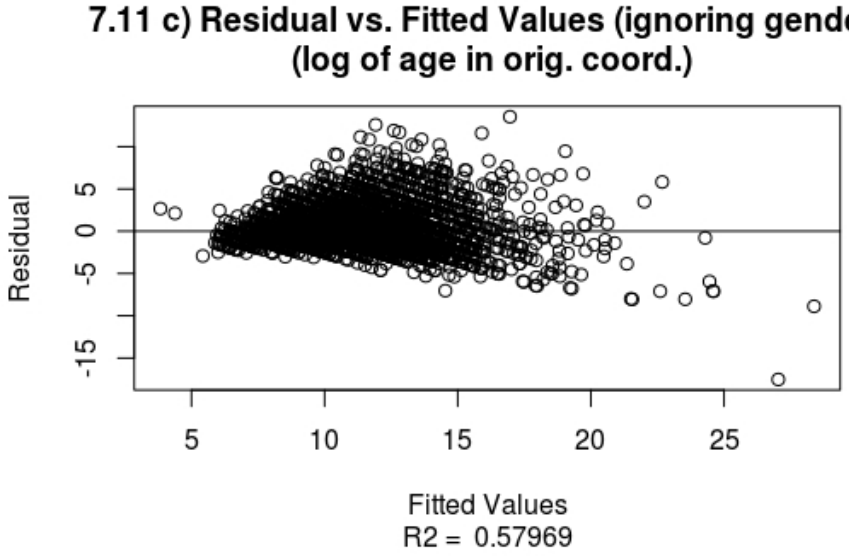
a) Linear regression predicting the age from measurements, ignoring gender. Residual vs. fitted values:



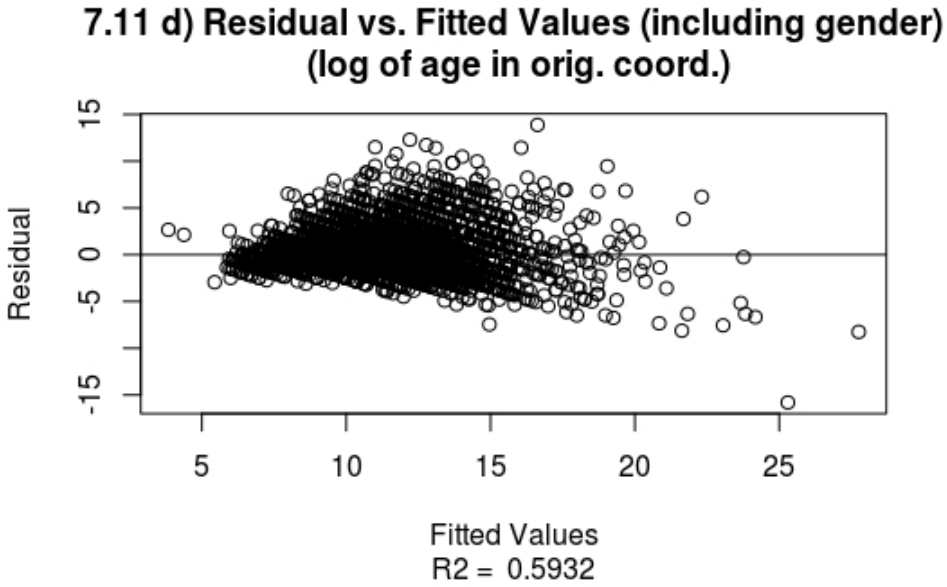
(b) Linear regression predicting the age from measurements, including gender. There are three levels for gender (three levels used: 1, 0, -1). Plot of residual vs. fitted values.



(c) Linear regression predicting the log of age from measurements, ignoring gender. Residual vs. fitted values.



(d) Linear regression predicting the log age from the measurements, including gender (same three levels). Residual vs. fitted values.



(e) We can evaluate regression results on the basis of evaluating R^2, looking at plots, looking for signs of a constant prediction, and looking at the randomness of the residual.
It is difficult to find significant differences among these four plots from the point of view of the residual randomness. There is some randomness, but also a lot of points are located in the center of the plot. So, one probably could use this model for prediction, but it may not be the best model for these data. The nature of the distribution of points on the plots is somewhat similar with some small differences. Therefore, it is hard to choose among these four models by just looking at the plot.
I would use an objective indicator which is R^2 in this case. It seems to be noticeably lower for the first group of plots (a and b) and higher for the second one (c and d). It is the largest in d) and equals

0.5932. So, one could say that the linear regression for the log of age that includes gender is the best one out of these 4 models because it has the largest R^2. And in general, the log models seem to perform slightly better in this exercise.

(f) In my code, I used glmnet to obtain plots of the cross-validated prediction error to see if a regularizer can improve the above four regressions. These are mean-squared error vs. log(lambda) plots showing various lambdas along with lambda.min and lambda.1se. The dotted vertical lines show a range of reasonable choices of lambda (from the lowest observed error to the error whose mean is within one standard error of the minimum). Based on the fact that I get very small values at these labmdas, I concluded that the regularization did not really help too much in this case because as a result the regularized version will not be that much different from the non-regularized version.
Here are the plots: