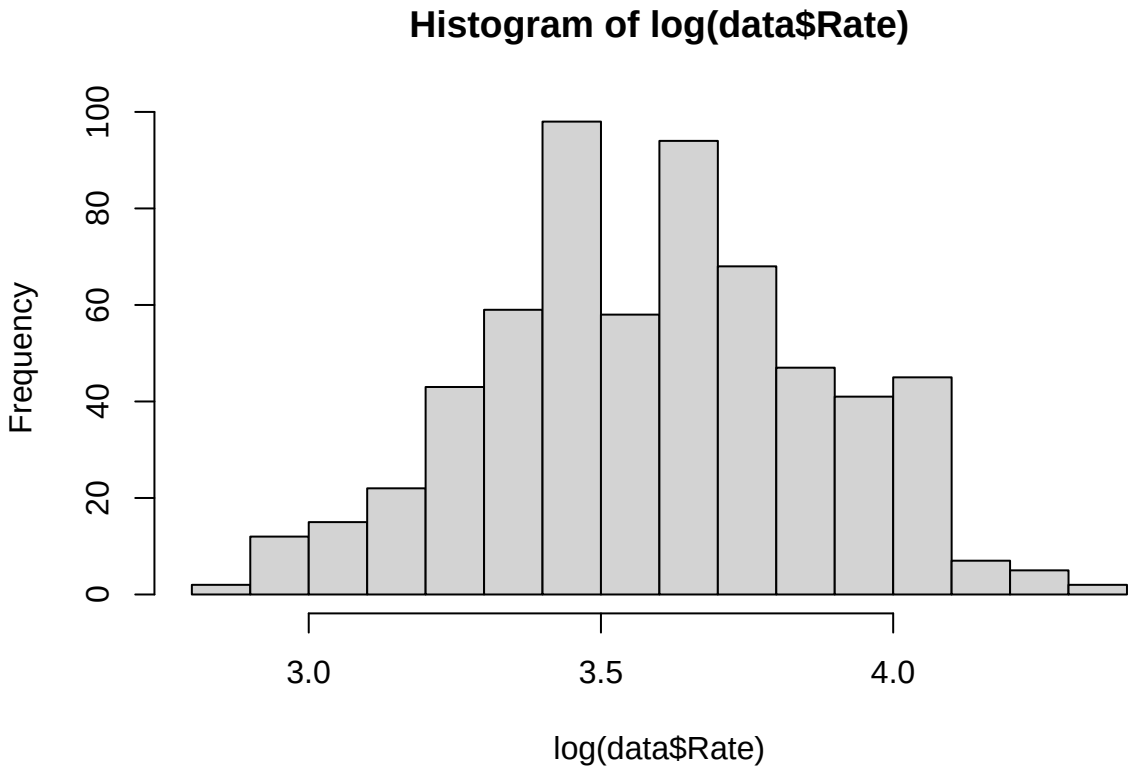
Question 1

(A)

Do exploratory analysis on the data and include a useful plot that a physician could use to assess a "normal" range of respiratory rates for children of any age between 0 and 3.

The dataset consists of 618 observations across 3 columns, X, Age, and Rate. The mean of the age and respiratory rate is 13.39 months and 37.74 respectively. There is a negative correlation between Rate and Age (-0.6903627). The histogram of Rate is a skewed distribution and hence taking log(Rate) could be use to assess a "normal" range of respiratory rates for children of any age between 0 and 3.



(B)

Write down a regression model for predicting respiratory rates from age. Make sure to use the right mathematical notation.

$\hat{y} = \hat{(}b_0) + \hat{(}b_1)$ * x

$log(\hat{Rate}) = 3.8451185 - 0.0190090 * Age$

(C)

Fit the model to the data and interpret your results.

The model has an intercept of 3.8451185. As this data is not centered, the meaning of this intercept is if the age of a child is 0 (unpractical), the respiratory rate will 3.84511. However, if the age is centered, the intercept indicates that the respiratory rate of a child of 13.39 months is 3.59058. The slope of Age is -0.0190090, which indicates that an increase of age by one month would decrease the respiratory rate by 0.0190090, keeping other variables constant. The percent of the variability in respiration rate explained by the regression model (R-squared) is 52.01%.

(D)

Include a table showing the output from the regression model including the estimated intercept, slope, residual standard error, and proportion of variation explained by the model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.845	0.01263	304.5	0
Age	-0.01901	0.0007357	-25.84	2.74e-100

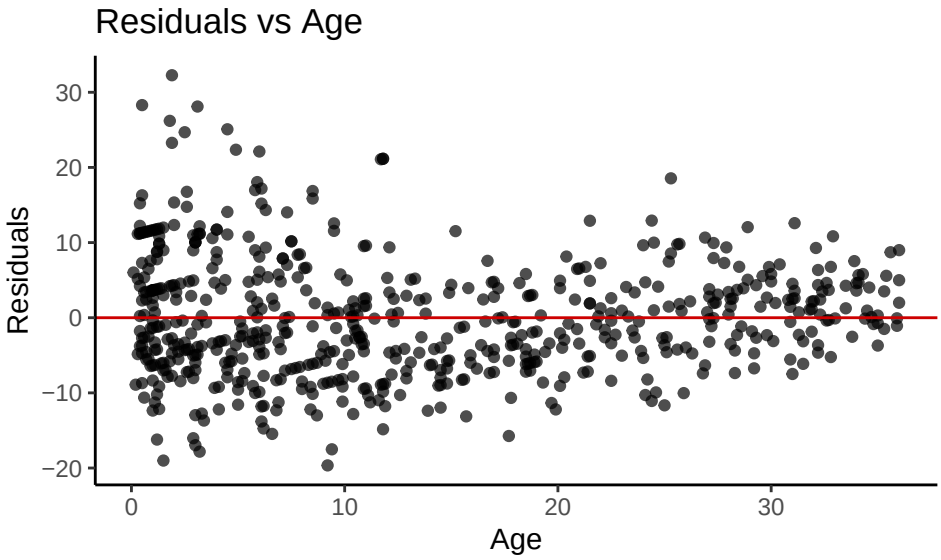
Table 2: Fitting linear model: log(Rate) ~ Age

Observations	Residual Std. Error	R^2	Adjusted R^2
618	0.1964	0.5201	0.5193

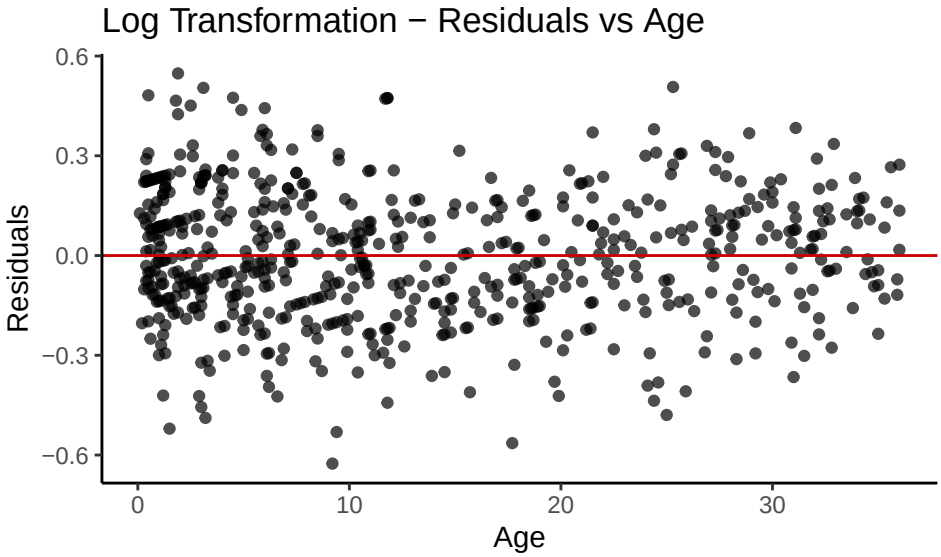
(E)

Is there enough evidence that the model assumptions are reasonable for this data? You should consider transformations (think log transformations, etc) if you think there's a violation of normality and/or linearity.

Testing for Linearity: To check whether the model (without log transformation) satisfies the linearity assumption, the Model Residual vs Age plot is shown below.

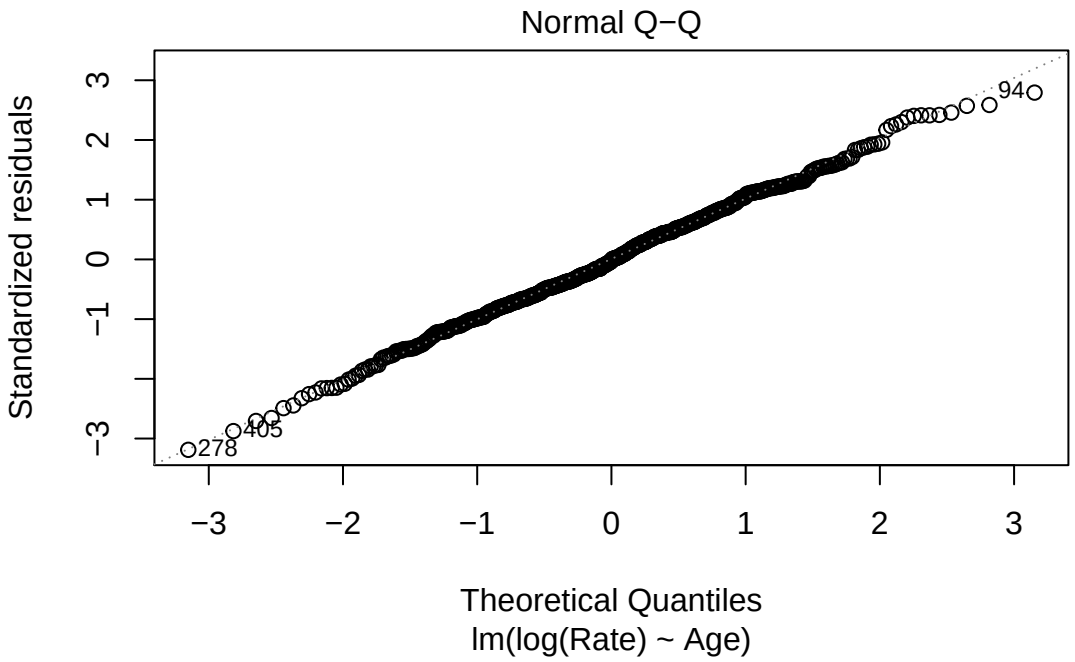


As shown above, there seems to be a quadratic curve pattern amongst data points in the plot. Hence, the log(Rate) is taken in consideration and the Model Residual vs Age plot (with log transformation) is plotted below.



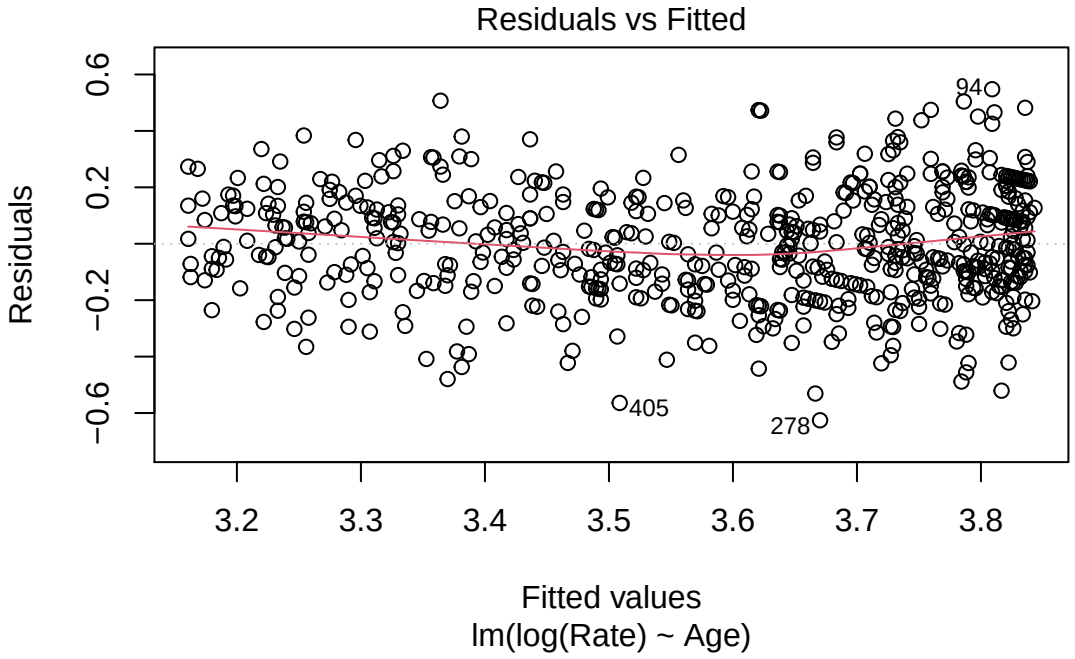
After transforming the y variable to its log, quadratic curve seems less evident and a higher degree of randomness can be observed. Hence, I used log(Rate) as the response variable for the regression model.

Testing for Normality: Using log(Rate) as the response, the Q-Q Plot is plotted below to identify Normality.



As the points lie on the 45 degree line, the normality assumption is not violated.

Testing for Independence and Equal Variance: Using log(Rate) as the response, the Residual vs Fitted Plot is plotted below to check if the independence and equal variance assumptions are not violated.



As the plots are spaced out throughout the X axis, and lie around the 0 residual mark on the Y axis, the

model is assumed to be independent and have equal variance.

(F)

Demonstrate the usefulness of the model by providing 95% prediction intervals for the rate for three individual children: a 1 month old, an 18 months old, and a 29 months old.

Age	fit	lwr	upr
1	45.88	31.18	67.53
18	33.21	22.58	48.86
29	26.95	18.31	39.67