

Question 4.6

- a. It is difficult to ultimately assess independence with the amount of information given. However, under the assumption that all 618 children were unique (no child was measured twice) then it seems that the data are at least roughly independent. There is some evidence for a slight non-linearity (Figure 1-Left). The residual plot suggests a heteroscedasticity (Figure 1-Left). The residuals do not appear to be normal (Anderson Darling $p < 0.00005$) and are right-skewed (Figure 1-Right). I did not test for outliers given the violations of the normality and homoscedasticity assumptions.

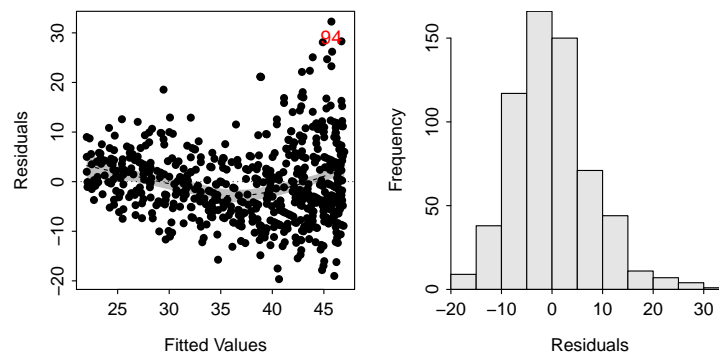


Figure 1. Residual plot (Left) and histogram of residuals (Right) for the simple linear regression of respiratory rate on child's age.

- b. The ratio of maximum to minimum age was 360 suggesting that the age variable could be transformed to logarithms. However, the trial-and-error method suggested leaving age untransformed and transforming respiratory rate to logarithms would provide an adequate fit. With only log-transformed respiratory rate there was no visual evidence for a non-linearity (Figure 2-Left), the residuals appeared homoscedastic (Figure 2-Left) and normal (Anderson-Darling $p = 0.3093$; Figure 2-Right), and there was no evidence for significant outliers (outlier test $p = 0.8544$). Thus, the assumptions appear to be adequately met on this transformed scale.

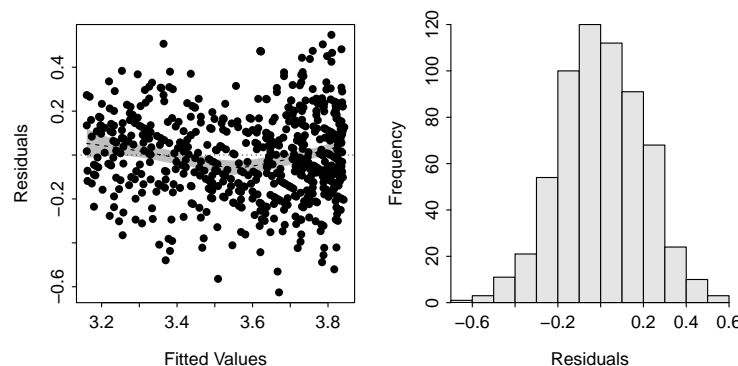


Figure 2. Residual plot (Left) and histogram of residuals (Right) from the simple linear regression of log-transformed respiratory rate on child's age.

- c. There is a significant relationship between the log respiratory rate and the age of a child ($p < 0.00005$; Table 1; Figure 3). Specifically, as the age of the child increases by one month then the average log respiratory rate *decreases* between 0.018 and 0.020.
- d. The predicted respiratory rate, corrected for back-transformation bias, for four different ages is shown in Table 2. For example, the predicted respiratory rate for a five month old child is between 29.5 and 63.8.

Table 1. ANOVA table for simple linear regression results of log-transformed respiratory rate on child's age.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
age	1	34496	34496	560.92	< 2.2e-16
Residuals	616	37884	61		
Total	617	72380			

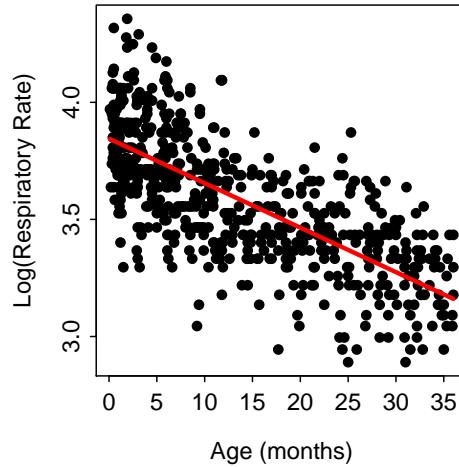


Figure 3. Fitted line plot for the simple linear regression of log-transformed respiratory rate on child's age.

Table 2. Respiratory rates for variously aged children predicted from the simple linear regression of log-transformed respiratory rate on child's age.

	obs	age	fit	lwr	upr
1	1	5	43.35232	29.46384	63.78746
2	2	6	42.53602	28.91028	62.58371
3	3	21	31.98339	21.73783	47.05793
4	4	22	31.38116	21.32759	46.17388

R Commands

```
> rusc <- read.csv("https://raw.githubusercontent.com/droglenc/NCDData/master/Rusconi.csv")
> lm1 <- lm(rate~age,data=rusc)
> adTest(lm1$residuals)
> residPlot(lm1,main="")
> hist(lm1$residuals,main="Residuals",main="")
> max(age)/min(age)
> transChooser(lm1)
> rusc$log.rate <- log(rusc$rate)
> lm2 <- lm(log.rate~age,data=rusc)
> fitPlot(lm2,xlab="Age (months)",ylab="Log(Respiratory Rate)",main="")
> residPlot(lm2,main="")
> hist(lm2$residuals,main="Residuals",main="")
> anova(lm2)
> summary(lm2)
> confint(lm2)
> nd <- data.frame(age=c(5,6,21,22))
> p.rate <- predict(lm2,nd,interval="prediction")
> exp(p.rate)*exp(anova(lm2)[2,3]/2)
```

Notes from Professor

-