Regression in R

Use USairpollution dataset

1) Basic linear model (lm)
   a) **Model1 <- lm(outcome.variable ~ predictor.variable, data = mydata)**
   b) Use **na.action=na.exclude** to exclude cases with missing data
   c) **summary(Model1)**
   d) Output
      i) Estimate - for (Intercept) this is the y-intercept or value of a, for the predictor variable this is the regression coefficient or slope or value of b
      ii) Std. Error - standard error of estimate for intercept and slope
      iii) t value - test values
      iv) Pr(>|t|) - p value
      v) Residual standard error - Standard error of estimate for the equation as a whole
      vi) Multiple R-squared - r^2 because there is only 1 predictor variable
   e) To get standardized coefficients use a function from the QuantPsyc package
      i) lm.beta(Model1)
      ii) p-value is the same as for the unstandardized coefficient

2) Confidence interval for intercept and b value
   a) confint(Model1)
   b) Output: lower and upper limits for intercept and slope

3) Plots to check assumptions
   a) **plot(Model1)**
      i) Residuals vs Fitted: residuals on Y axis and fitted values on x axis; points should not have any structure or pattern; and scatter should not increase as the fitted values get bigger
      ii) Normal Q-Q: points should be along the line, which means the errors are normally distributed, points not on the line indicate outliers in the data
      iii) Scale-Location: similar to Residuals vs Fitted, but uses the square root of the standardized residuals; points should be along the line
      iv) Residuals vs Leverage: use to see which Y values have the biggest effects on the parameter estimates
   b) Histogram to check for normality of the distribution of residuals
      i) First, compute Studentized residuals, which are like standardized residuals but they ignore the current data point
      sresid <- studres(Model1)
      ii) Then, create a histogram with a normal curve

```
hist(sresid, freq=FALSE,
     main="Distribution of Studentized Residuals")
xfit<-seq(min(sresid),max(sresid),length=40)
yfit<-dnorm(xfit)
lines(xfit, yfit)
```