**Pollution and Mortality with Regression**

a) For this exercise, we will use a data set that describes properties of 60 Standard Metropolitan Statistical Areas (a standard Census Bureau designation of the region around a city) in the United States. These are collected from a variety of sources, focusing on mortality and pollution.

First, look at histograms of mortality and NOx (nitrous oxide concentration). Given that NOx has a lot of variability, compute another column with the logarithm of NOx.

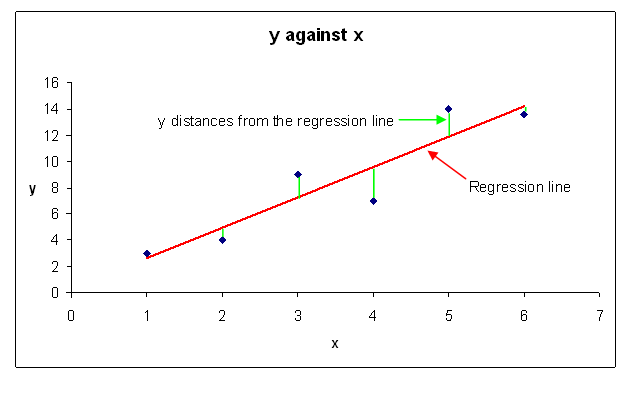
In R, use the log function to calculate the appropriate value:

log()

Show a scatter plot with mortality plotted against log(NOx)

What is the relationship between these two variables?

We will draw a best-fit line to represent this relationship using simple linear regression analysis. To do this, use simple linear regression analysis with mortality (dependent variable) and log(NOx) (independent variable). Recall: simple linear regression of two variables finds the line which minimizes the square errors.



When you run a regression of *Y* on *X*, you are assuming the following model:

*Y* = *b*0 + *b*1*X* + ε

ε ~ N(0, *σ*2)

Where *b*1 and *b*0 are constants (the slope and intercept of the line)

And ε is a normally distributed random variable representing the residuals or error in the model.

Regression helps us estimate *b*1 and *b*0, giving us the best-fit line.

So, if *Y* is mortality and *X* is log(NOx), what are your estimates of *b*1 and *b*0 (the intercept and slope of the regression)?

b­1 = 15.1 (slope)

b0 = 905.6 (intercept)

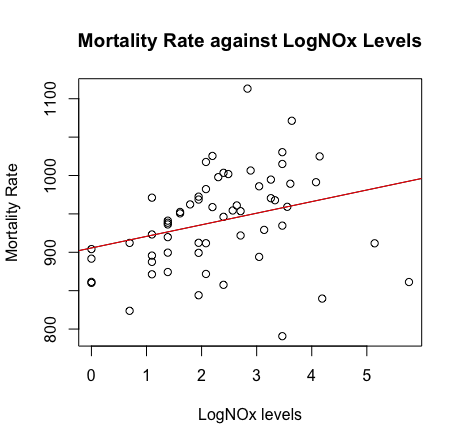
What is the relationship between mortality and the log of nitrous oxide concentrations across US cities?

i) Provide the slope, *y*-intercept and full formula below.

b­1 = 15.1 (slope)

b0 = 905.6 (intercept)

 ŷ  = 15.1 X + 905.6

ii) Paste your clearly labeled plot below

iii) Note why mortality is the dependent variable.

Mortality is the dependent level because we are insinuating that the nitrous oxide levels could be a potential contributor to the mortality rates.

It would not make sense if the nitrous oxide levels were the dependent variable and mortality rates were the independent variable.

**PAUSE HERE, CHECK ANSWERS WITH CLASS, BRING FORWARD ANY QUESTIONS**

b) Please check that your results from above match the mathematical derivation by manually calculating the slope and *y-*intercept for simple linear regression analysis in R. Type your resulting values for the slope and *y*-intercept below:

sYY = 228398.3

sXX = 87.25589

sXY = 1317.474

slope = sXY / sYY = 905.61

intercept = mean(x) - slope\* mean(y) = 15.098

c) Use your slope and *y*-intercept values to generate a column for “predicted mortality” based on your estimated linear regression equation. Also, calculate another column for the residuals.

Take the first city: Akron OH.

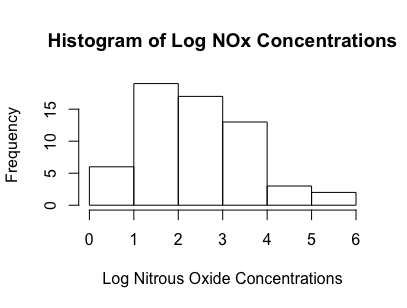
According to your model what was the residual term for that city εAKRON?

Residual for Akron = -24.63192384

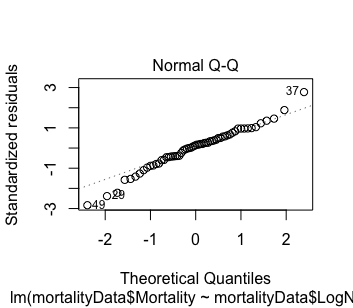
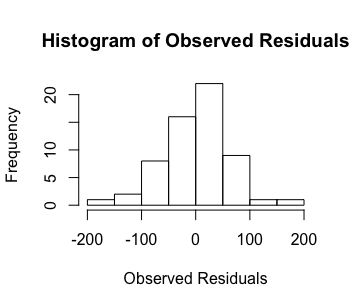
**PAUSE HERE, CHECK ANSWERS WITH CLASS, BRING FORWARD ANY QUESTIONS**

d) Perform residual/normality analysis on these results.

1) visually test the normality of the log(NOx) variable

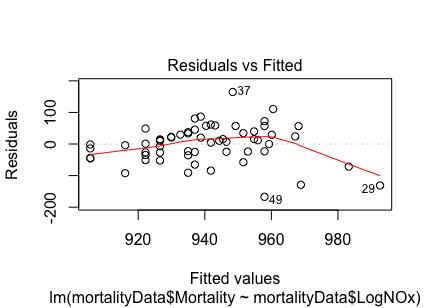
Yes

2) visually test the normality of the residuals

 Yes on the observed

3) visually test the constant variance assumption

i) Clearly label your plots and paste them below.

observed residuals

ii) Was each assumption met? (1-3 sentences)

Independent variable is normally distributed

Residuals for the observed was normally distributed

Residuals for the observed was mostly homoscedastic

e) What percentage of the variation in mortality was explained by log(NOx)? (Recall the coefficient of determination)

0.087 = 8.7%

**PAUSE HERE, CHECK ANSWERS WITH CLASS, BRING FORWARD ANY QUESTIONS**