Government statisticians in England conducted a study of the relationship between smoking and lung cancer. The data concern 25 occupational groups and are condensed from data on thousands of individual men. The explanatory variable is the number of cigarettes smoked per day by men in each occupation relative to the number smoked by all men of the same age. This smoking ratio is 100 if men in an occupation are exactly average in their smoking, it is below 100 if they smoke less than average, and above 100 if they smoke more than average. The response variable is the standardized mortality ratio for deaths from lung cancer. It is also measured relative to the entire population of men of the same ages as those studied, and is greater or less than 100 when there are more or fewer deaths from lung cancer than would be expected based on the experience of all English men.

1. Plot the data in the file smoke.txt. The first variable is the smoking index smoke and the second is the mortality index mort. An appropriate graph would be a scatter plot to explore the data. Which variable should go on the x-axis and which on the y-axis? Describe any patterns that you observe. Does a linear relationship between smoke and mort seem plausible?

2. Many of you have probably studied simple linear regression, which is a method used to fit a straight line model to data in order to describe the relationship between a response variable Y and an explanatory variable X. In simple linear regression, we assume that we can model the relationship between the ith observed values of Y and X as follows:

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \]

where \( \beta_0 \) is the intercept of the line, \( \beta_1 \) is the slope, and \( i = 1, 2, \ldots, n \). The term \( \epsilon_i \) in the model is an “error” term that expresses the random deviation of the observed \( Y_i \) from the value of the true regression line \( \beta_0 + \beta_1 X_i \).
The above statement is not a complete description of the model. The complete description also includes some important statistical assumptions:

- $\epsilon_i$ and $\epsilon_j$ are independent if $i \neq j$
- The error terms $\epsilon_i$ are normally distributed with mean 0 and variance $\sigma^2$.
- although the model explicitly allows for measurement error in the $Y$ variable, measurements made on $X_i$ are known precisely.

One of the commonly used methods of fitting a straight line to data is called linear regression, or least squares.

Draw a picture that illustrates the principle behind this method.
3. Using the method of least squares, the formulas for estimating the intercept $\beta_0$ and the slope $\beta_1$ are as follows:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

and

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X},$$

where $\bar{X}$ and $\bar{Y}$ are the means of the $X$ and $Y$ observations, respectively. Since these estimates are functions of data, they will change from data set to data set—that is, $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables. Furthermore, the standard errors for $\hat{\beta}_0$ and $\hat{\beta}_1$ can be computed as follows:

$$\text{SE}[\hat{\beta}_0] = \text{MSE} \left[ \frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right]$$

and

$$\text{SE}[\hat{\beta}_1] = \frac{\text{MSE}}{\sum_{i=1}^{n}(X_i - \bar{X})^2},$$

where the mean square error $\text{MSE} = \frac{\sum_{i=1}^{n}e_i^2}{n-2}$ and $e_i$ is the $i$th residual.

These formulas look nasty and as you can imagine, they are even worse when extended to the case where our model has more than one explanatory variable. As a result, we often frame the estimation problem in terms of linear algebra.

Write down the matrix expressions for estimating the intercept and slope, as well as their covariance matrix.
4. Now let’s apply these formulas to the data set smoke.txt. You can write your own code or use the code in smoke.m. Fit the model with the intercept term. What do you conclude about the need to include an intercept term in the model?

5. Fit the model without an intercept term. What is your estimate of the Lung cancer mortality rate when the smoking ratio is 100 (exactly average in their smoking) in an occupation?

6. No statistical analysis is complete without a complete check of the model assumptions that were given previously. Use the plots provided in the MATLAB code smoke.m (or any other methods you can think of) to test the model assumptions.
7. Now repeat your analysis on the second data set hubble.txt. These data were collected by Hubble. And details are on “http://lib.stat.cmu.edu/DASL/Stories/Hubble'sConstant.html”. Remember to follow the important steps in any statistical analysis, namely:

- plot the data
- propose a model
- fit the model (this includes standard error estimates and/or confidence intervals for parameter estimates) and
- check the fit of the model, as well as the other model assumptions.

Does this data support the Big Bang Theory? Based on this second data set, estimate the age of the universe. How does your estimate compare with the currently held belief that the universe is between 10 and 15 billion years old?