APPLIED STATISTICS
TUTORIAL 2

The data files for this tutorial can be found on Wattle. The questions for this tutorial have been revised directly from the class text "The Statistical Sleuth".

Question 1 (revised based on ex 7.27 from Statistical Sleuth)
The file "ex0727.csv" contains measured distances and recession velocities for 10 clusters of nebulae. According to a theory by Hubble the mean of the measured distance, as a function of velocity, should be β_1*velocity (i.e., μ(distance|velocity) = β_1*velocity), and β_1 is the age of the universe.

b) Using Hubble's theory what is the estimated age of the universe? (Hint: The R function lm() includes an intercept by default. lm(Y~X-1) fits the SLR without an intercept, i.e., $\mu(Y|X)= \beta_1 X$.

Question 2 (revised based on ex 7.29 from Statistical Sleuth)
Black wheatears are small birds of Spain and Morocco. Males of the species demonstrate an exaggerated sexual display by carrying many heavy stones to nesting cavities. Different males carry somewhat different sized stones, prompting a study of whether larger stones may be a signal of higher health status. A study was conducted (M. Soler et al.) which calculated the average stone mass (g) carried by each of 21 male wheatears, along with T-cell response measurements reflecting their immune systems' strengths. The file "ex0729.csv" contains the data.

c) For wheatears that carry stones with an average mass of 2g, what would you estimate their mean T-cell response to be? Comment on this estimate.

Question 3 (revised based on ex 8.18 from Statistical Sleuth)
One of the most dangerous contaminants deposited over European countries following the Chernobyl accident of April 1987 was radioactive cesium. To study cesium transfer from contaminated soil to plants, researchers collected soil samples and samples of mushroom mycelia from 17 wooded locations in Umbria, Central Italy from August 1986 to November 1989. Measured concentrations of cesium (Bq/Kg) in the soil and in the mushrooms are contained in the file "ex0818.csv".

a) Construct a scatterplot of Y = concentration in mushrooms and X = concentration in soil. What do you notice?
b) Fit a simple linear regression using Y = concentration in mushrooms and X = concentration in soil. Produce a plot of the fitted regression line superimposed on the scatterplot of points.
c) Repeat part (b) with point 17 removed. What do you notice?

Question 4 (Simple Linear Regression)

Consider the simple linear regression model $\mu\{Y|X\} = \beta_0 + \beta_1 X$ for the observations $\{Y_i, X_i\}_{i=1}^n$, and the least squares estimates $\hat{\beta}_0, \hat{\beta}_1$ for the coefficients β_0, β_1 can be obtained. Based on the formulas given in Lecture Notes 1, the fitted values $\hat{Y}_i$ and residuals res_i can be given for $i = 1, \cdots, n$.

Part 1. The sample mean of the residuals can be defined as

$$\overline{\text{res}} := \frac{1}{n}\sum_{i=1}^{n} \text{res}_i.$$

Show that

$$\overline{\text{res}} = 0.$$

(Hint: use the formula on pages 16-17 of Lecture Notes 1.) Based on this result, show that the sample variance of the residuals is

$$s^2_{\text{res}} = \frac{1}{n-1}\sum_{i=1}^{n} (\text{res}_i - \overline{\text{res}})^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2.$$

Part 2. The sample mean of the fitted values can be defined as

$$\bar{\hat{Y}} := \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i.$$

Show that

$$\bar{\hat{Y}} = \bar{Y},$$

where $\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$ is the sample mean of response. Based on this result, show that the sample variance of the fitted values is

$$s^2_{\hat{Y}} = \frac{1}{n-1}\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{\hat{Y}}\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2.$$

Part 3. By using R, please follow the steps below to interpret the sampling distribution of the estimated mean of response $\hat{\mu}\{Y|X\} = \hat{\beta}_0 + \hat{\beta}_1 X$. Please attach the R codes and R output, as well as the interpretation. (Hint: similar to the codes on pages 7-11 of Lecture Notes 2.)

STEP 1: Specify $\beta_0 = 1$ and $\beta_1 = 0$.

STEP 2: Suppose the observations of X are $1, 2, \cdots, 100$, so the number of observations $n = 100$.

STEP 3: Generate $\mathcal{E}_1, \cdots, \mathcal{E}_n$ from the standard normal distribution [$N(0,1)$ with mean 0 and variance 1].

STEP 4: Generate $Y_i = \mu(Y_i|X_i) + \mathcal{E}_i$, $i = 1, \cdots, n$.

STEP 5: Obtain the least squares estimates $\hat{\beta}_0, \hat{\beta}_1$ and the estimated mean of response given $X = 2.5$.

STEP 6: Repeat Step 3 – Step 5 1000 times and obtain 1000 different estimated values of the mean of response.

STEP 7: Draw the histogram of the 1000 different estimated values of the mean of response and compute the sample mean and sample standard deviation of those values.