Unit 9 – Regression and Correlation

Homework #14 (Unit 9 – Regression and Correlation)

Due: Friday December 11, 2015
Last submission date for credit: Friday December 11, 2015

Consider the following study of the relationship of cigarette consumption and lung cancer. The following are data from Sir Richard Doll's 1955 study. There are 11 paired observations (X,Y). X = per capita cigarette consumption (the year is 1930). Y = the number of lung cancer cases per 100,000 (the year is 1950). Each observation is from a different country.

Country	X = cigarette consumption (per capita in 1930)	Y = lung cancer cases (per 100,000 in 1950)
USA	1300	20
Great Britain	1100	46
Finland	1100	35
Switzerland	510	25
Canada	500	15
Holland	490	24
Australia	480	18
Denmark	380	17
Sweden	300	11
Norway	250	9
Iceland	230	6

1. The first step in the analysis is to look at a scatterplot of the data. By any means you like (by hand is just fine), construct an XY scatterplot of these data.

2. Interpret the graph you produced in exercise #1 with respect to form, direction, and strength.

3. By hand, or using Excel, or using any software you like, calculate the values of the following:

   a) $\bar{X}$
   b) $\bar{Y}$
   c) $S_{XY} = \sum_{i=1}^{11} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$
   d) $S_{XX} = \sum_{i=1}^{11} \left(X_i - \bar{X}\right)^2$
   e) $S_{YY} = \sum_{i=1}^{11} \left(Y_i - \bar{Y}\right)^2$

4. Now you have what you need to solve for the least squares estimate of the slope and intercept. By hand, or using Excel, or using any software you like, calculate the values of the following:

   a) Estimated slope, $\hat{\beta}_1 = \left[\dfrac{\sum_{i=1}^{11} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{11} \left(X_i - \bar{X}\right)^2} \right] = \left[\dfrac{S_{XY}}{S_{XX}} \right]$

   b) Estimated intercept, $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

5. State the fitted line and interpret it.

6. By hand, or using Excel, or using any software you like, calculate the values of the following sums of squares that are in the analysis of variance:

a) Total sum of squares, corrected = SST $=\sum_{i=1}^{11}\left(Y_i-\bar{Y}\right)^2$

hint – This is the same as S_{YY} in #3

b) Regression sum of squares = SSR = $\sum_{i=1}^{11}\left(\hat{Y}_i-\bar{Y}\right)^2=\hat{\beta}_1^2\sum_{i=1}^{11}\left(X_i-\bar{X}\right)^2$

hint – Of the two formulae shown, the right hand formula will be easier to do by hand!

c) Error sum of squares = SSE = $\sum_{i=1}^{11}\left(Y_i-\hat{Y}\right)^2=SST-SSR$

hint – Of the two formulae shown, the right hand formula will be easier to do by hand!

7. Complete the following analysis of variance table by supplying the numeric values of the df, sums of squares, mean squares and F statistic.

Source	df	Sum of Squares	Mean Square	F-Statistic
Regression	1	$SSR=\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2$	SSR/1	
Error	(n-2)	$SSE=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2$	SSE/(n-2)	
Total, corrected	(n-1)	$SST=\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2$		

Tip! – Mean square = (Sum of squares)/(degrees of freedom,df)

8. Perform and interpret the overall F test.