Problem set 1
Smoking and Lung Cancer
DUE: Wed. August 26, 2014

Observation No.	Country	Cigarettes consumed per capita in 1930 (X)	Lung cancer deaths per million people in 1950 (Y)
1	Switzerland	530	250
2	Finland	1115	350
3	Great Britain	1145	465
4	Canada	510	150
5	Denmark	380	165

1 Question 1

a) The sample means of X and Y, $\overline{X}$ and $\overline{Y}$.

$$\overline{X} = \frac{\sum\limits_{n=1}^{n} X_i}{n} = 736$$

$$\overline{Y} = \frac{\sum\limits_{n=1}^{n} Y_i}{n} = 276$$

b) The standard deviations of X and Y, s_X and s_Y.

$$s_X = \sqrt{\frac{\sum\limits_{n=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}} = 364.4071$$

$$s_Y = \sqrt{\frac{\sum\limits_{n=1}^{n} \left(Y_i - \overline{Y}\right)^2}{n-1}} = 132.3537$$

c) The correlation coefficient, r, between X and Y.

$$r = Cov(X,Y) = \frac{\sum\limits_{n=1}^{n} (X - \overline{X})(Y - \overline{Y})}{s_X s_Y (n-1)} = 0.9262529$$

d) $\hat{\beta}_1$, the OLS estimated slope coefficient from the regression $Y_i = \beta_0 + \beta_1 X_i + u_i$.

$$\hat{\beta}_1 = \frac{\sum\limits_{n=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum\limits_{n=1}^{n} \left(X_i - \overline{X}\right)^2} = 0.3364177$$

e) $\hat{\beta}_1$, the OLS estimated intercept term from the same regression.

$$\beta_0 = \overline{Y} - \beta_1 \overline{X} = 28.39656$$

f) $\hat{Y}_i$, $i = 1, \ldots, n$, the predicted values for each country from the regression.

$$Y_i = \beta_0 + \beta_1 X_i$$

$\hat{Y}_i$	207	404	414	200	156

g) μ_i, the OLS residual for each country.

$$\mu_i = Y_i - \hat{Y}_i$$

μ_i	43.3	-53.5	51.41	-49.97	8.76

2 Question 2

Now calculate the statistics in question 1 using STATA. On the STATA output file, find and label the items in Question 1.
At this section I have used R. You can see code at the attached file. The script output is:

[1] $\overline{X} = 736$, $\overline{Y} = 276$
[2] $s_x = 364.4071$, $s_y = 132.3537$
[3] r = 0.9263
[4] $\hat{\beta}_1$ and $\hat{\beta}_0$
Call: $lm(Y \sim X)$
Coefficients:

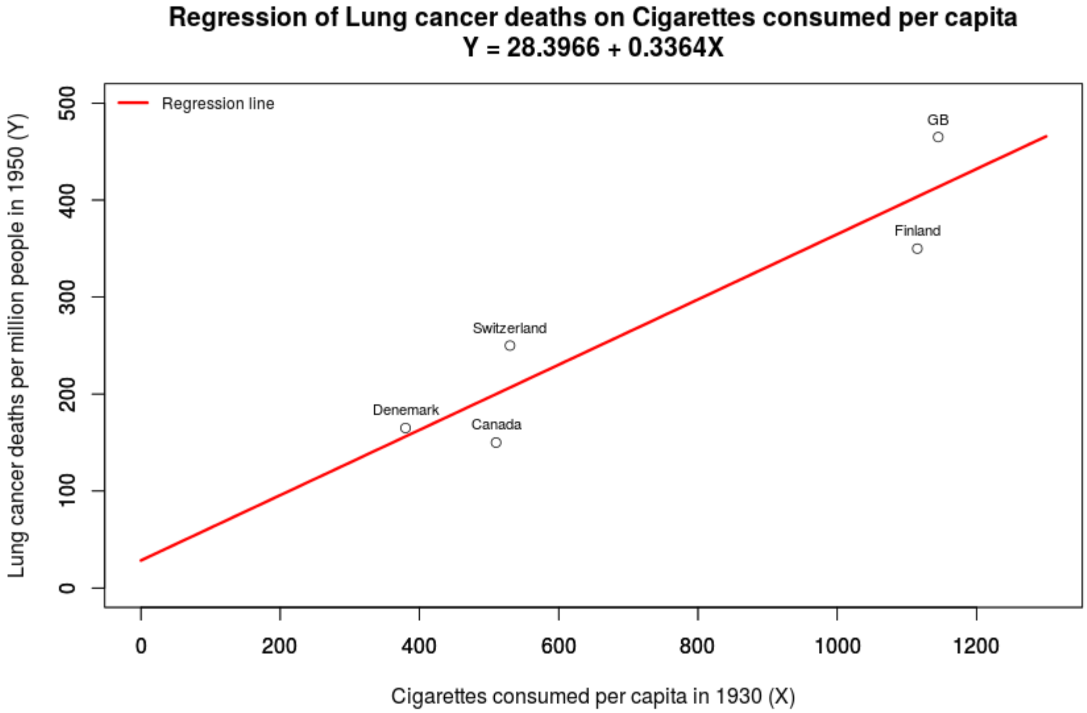
Intercept	data$X$
28.3966	0.3364

[5] $\hat{Y}_i$
Call: $predict(lm(Y \sim X))$

206.6979	403.5023	413.5948	199.9696	156.2353

3 Question 3

On graph paper or using a spreadsheet, graph the scatterplot of the five data points and the regression line. Be sure to label the axes, the data points, the residuals, and the slope and intercept of the regression line. (If you know the graphing commands in STATA you may do this using STATA. STATA has good graphing features but the commands are a little complicated and there is no need to learn them now – they will be covered later in the course.)


Regression of Lung cancer deaths on Cigarettes consumed per capita
Y = 28.3966 + 0.3364X
Regression line
GB
Finland
Switzerland
Denemark
Canada
Lung cancer deaths per million people in 1950 (Y)
Cigarettes consumed per capita in 1930 (X)