Unit 9 – Regression and Correlation
Homework #14 (Unit 9 – Regression and Correlation)

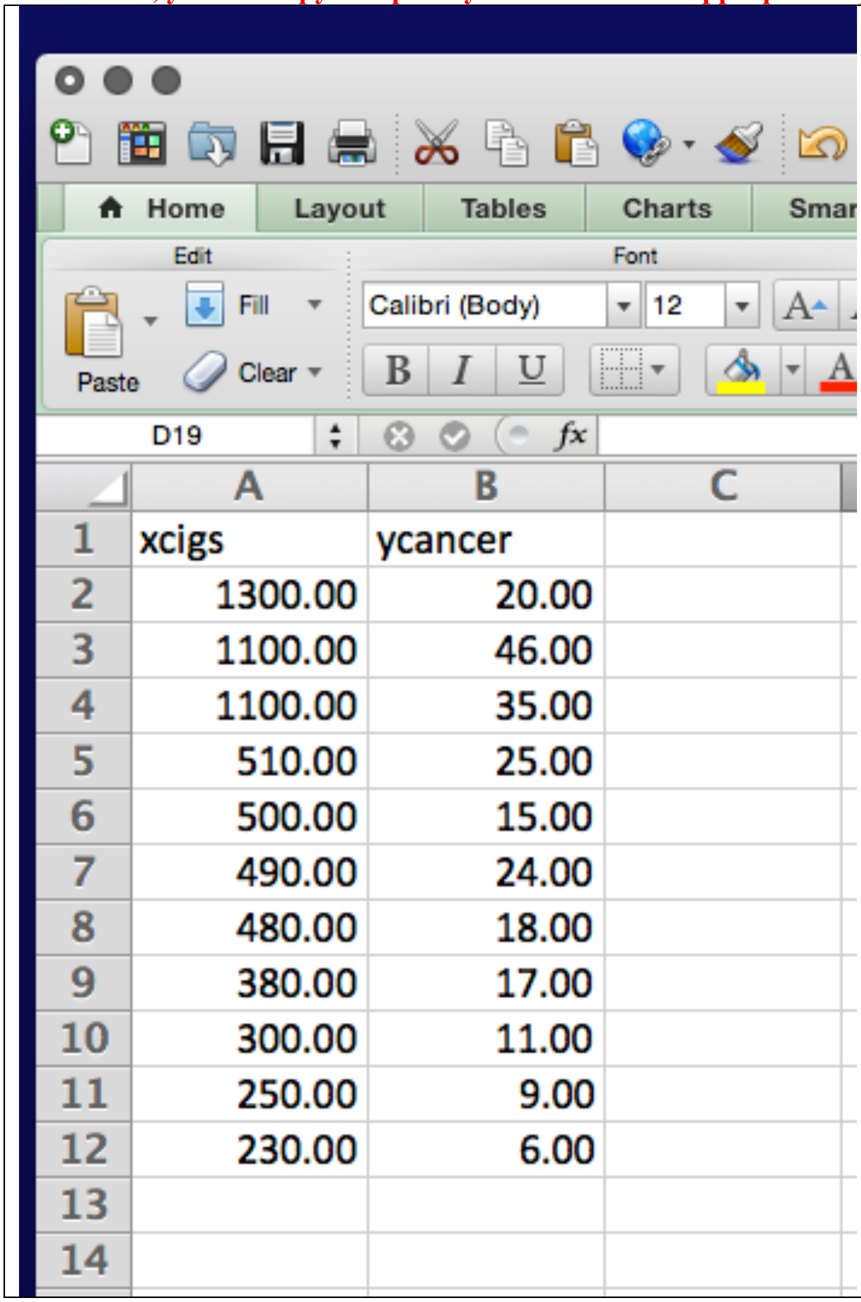
SOLUTIONS

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.

Suggested Preliminary – Create a little excel data set, taking care to format your columns of data as numeric. From here, you can copy and paste your data into an appropriate application.

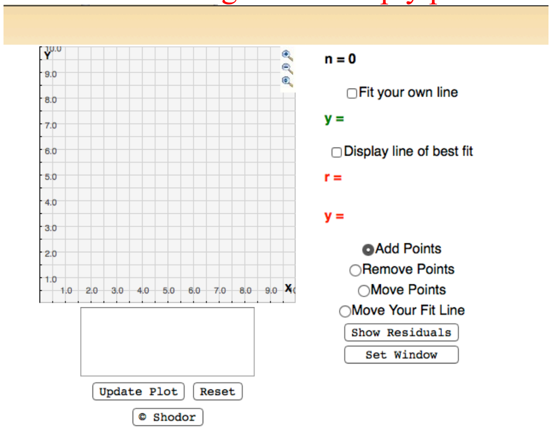


	A	B	C
1	xcigs	ycancer	
2	1300.00	20.00	
3	1100.00	46.00	
4	1100.00	35.00	
5	510.00	25.00	
6	500.00	15.00	
7	490.00	24.00	
8	480.00	18.00	
9	380.00	17.00	
10	300.00	11.00	
11	250.00	9.00	
12	230.00	6.00	
13			
14			

Scatterplot Using the Shodor Applet for Regression

Launch http://www.shodor.org/interactivate/activities/Regression/

You will be brought to an empty plot with an empty data set just below



n = 0

□ Fit your own line

y =

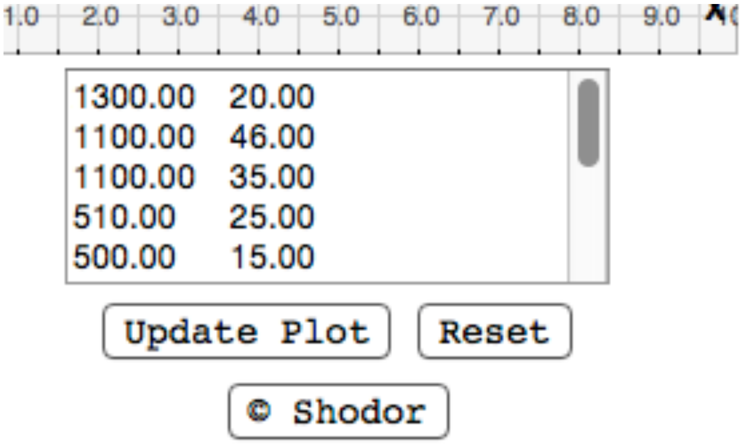
□ Display line of best fit

r =

y =

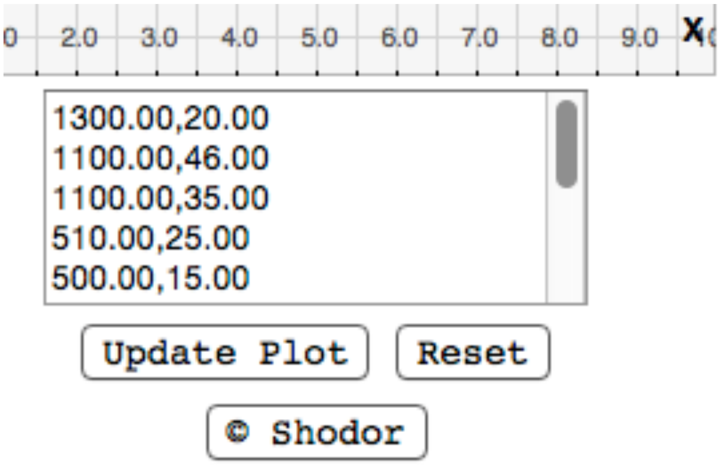
● Add Points
○ Remove Points
○ Move Points
○ Move Your Fit Line

Copy and paste your X-Y data. Notice that the X and Y data are separated by a blank. Don't click on "update plot" yet. Shodor doesn't like spaces between X and Y.



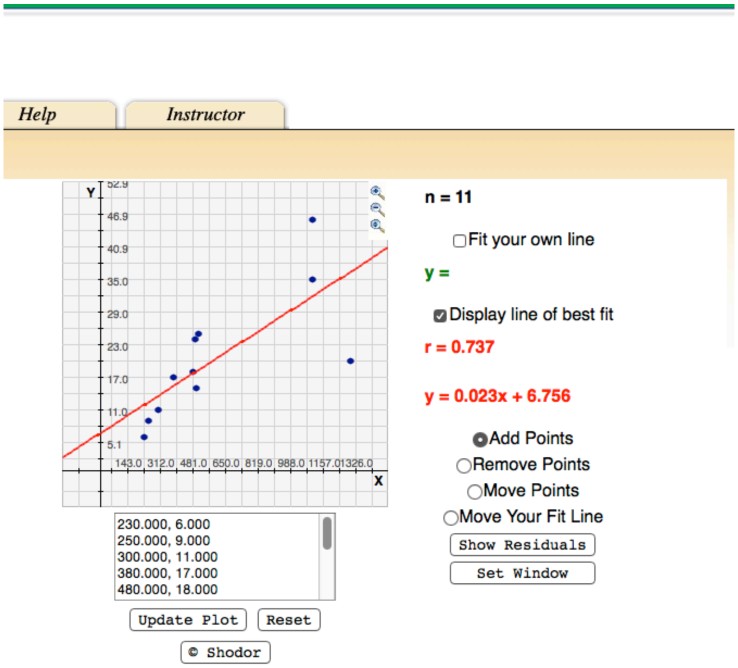
```
1300.00   20.00
1100.00   46.00
1100.00   35.00
510.00    25.00
500.00    15.00
```

Insert a comma between each X and Y as shown below. Be sure you scroll down and edit every row!



```
1300.00,20.00
1100.00,46.00
1100.00,35.00
510.00,25.00
500.00,15.00
```

Click update plot. If you like, click also on line of best fit. You should now see:



n = 11

□ Fit your own line

y =

☑ Display line of best fit

r = 0.737

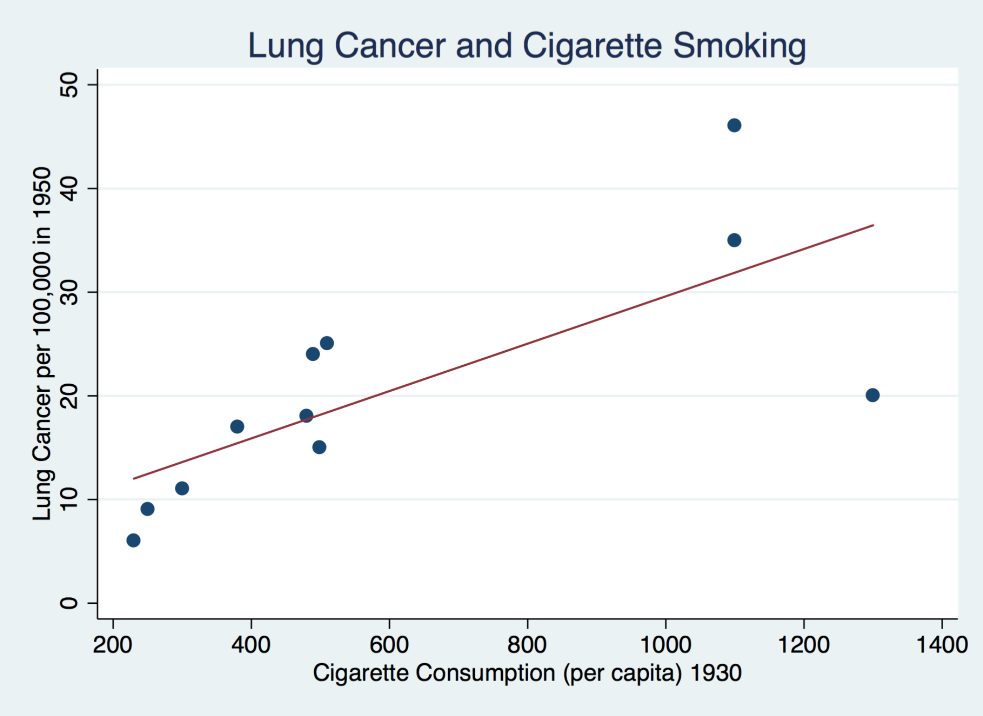
y = 0.023x + 6.756

● Add Points

```
230.000, 6.000
250.000, 9.000
300.000, 11.000
380.000, 17.000
480.000, 18.000
```

Scatterplot Using Stata

```
. *  Initialize data set
. generate xcigs=.
. generate ycancer=.

. *(2 variables, 11 observations pasted into data editor)

. label variable xcigs "Cigarette Consumption (per capita) 1930"
. label variable ycancer "Lung Cancer Cases (per 100,000) 1950"

. * Preliminary - Get min and max of each of X and y for setting the axes

. tabstat xcigs, statistics(min max)

    variable |       min       max
-------------+--------------------
       xcigs |       230      1300
----------------------------------

. tabstat ycancer, statistics(min max)

    variable |       min       max
-------------+--------------------
     ycancer |         6        46
----------------------------------

. graph twoway (scatter ycancer xcigs) (lfit ycancer xcigs), xlabel(200(200)1400) ylabel(0(10)50)
ytitle("Lung Cancer per 100,000 in 1950") title(“Lung Cancer and Cigarette Smoking") legend(off)
```



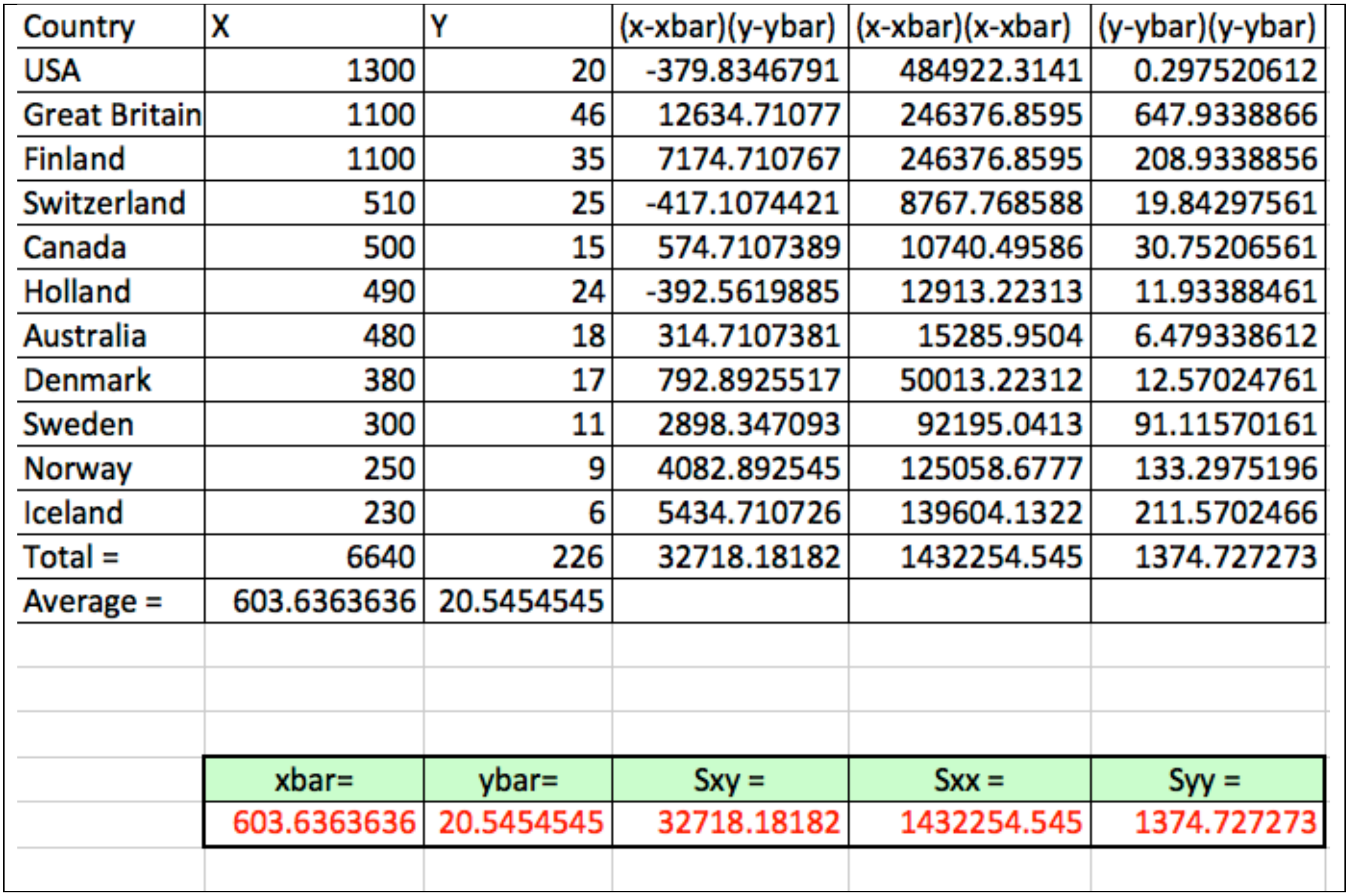
2. Interpret the graph you produced in exercise #1 with respect to form, direction, and strength.

This scatter suggests a linear relationship between cigarette consumption (X) and lung cancer cases (Y) that is positive, with higher cigarette consumption being associated with higher numbers of cancer cases. There are no outliers. However, there are more data in the lower left quadrant of this plot; thus, the full nature and strength of the association may be difficult to assess.

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

   a) $\bar{X}$ **= 603.6363636**
   b) $\bar{Y}$ **= 20.5454545**
   c) $S_{XY} = \sum_{i=1}^{11}(X_i - \bar{X})(Y_i - \bar{Y})$ **= 32718.18182**
   d) $S_{XX} = \sum_{i=1}^{11}(X_i - \bar{X})^2$ **= 1432254.545**
   e) $S_{YY} = \sum_{i=1}^{11}(Y_i - \bar{Y})^2$ **= 1374.727273**

Excel Worksheet: next page….



Country	X	Y	(x-xbar)(y-ybar)	(x-xbar)(x-xbar)	(y-ybar)(y-ybar)
USA	1300	20	-379.8346791	484922.3141	0.297520612
Great Britain	1100	46	12634.71077	246376.8595	647.9338866
Finland	1100	35	7174.710767	246376.8595	208.9338856
Switzerland	510	25	-417.1074421	8767.768588	19.84297561
Canada	500	15	574.7107389	10740.49586	30.75206561
Holland	490	24	-392.5619885	12913.22313	11.93388461
Australia	480	18	314.7107381	15285.9504	6.479338612
Denmark	380	17	792.8925517	50013.22312	12.57024761
Sweden	300	11	2898.347093	92195.0413	91.11570161
Norway	250	9	4082.892545	125058.6777	133.2975196
Iceland	230	6	5434.710726	139604.1322	211.5702466
Total =	6640	226	32718.18182	1432254.545	1374.727273
Average =	603.6363636	20.5454545			

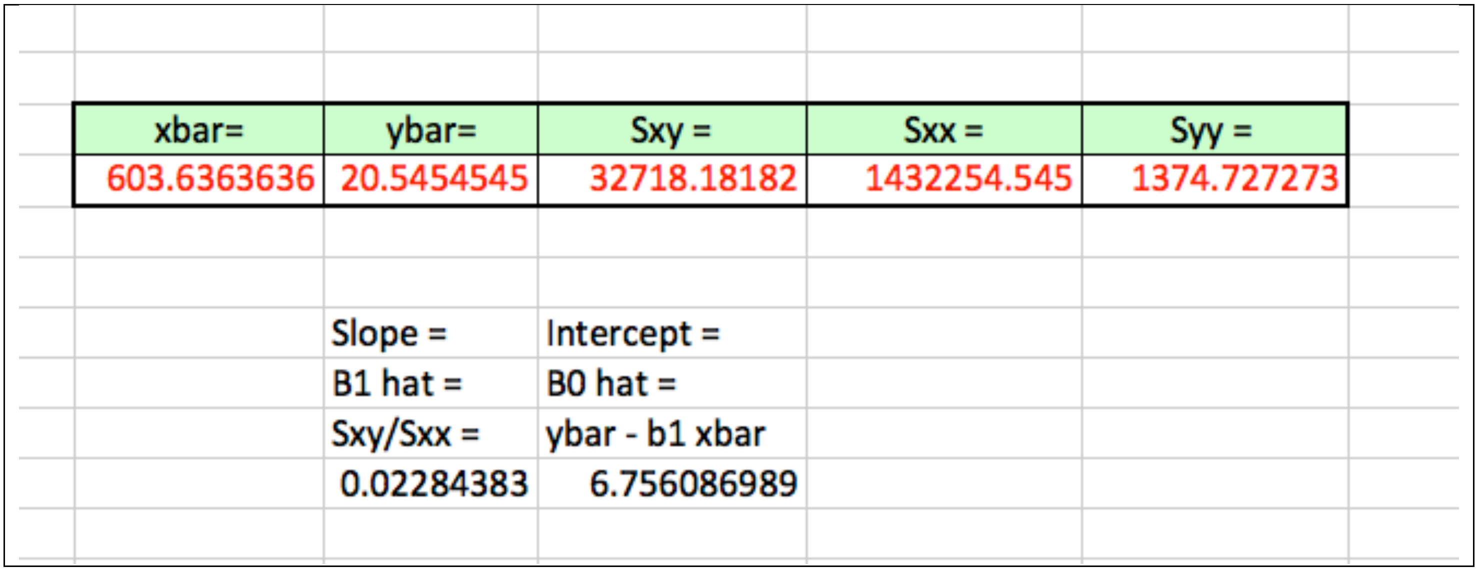
xbar=	ybar=	Sxy =	Sxx =	Syy =
603.6363636	20.5454545	32718.18182	1432254.545	1374.727273

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, $\hat{\beta}_1 = \left[\frac{\sum_{i=1}^{11}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{11}(X_i - \bar{X})^2} \right] = \left[\frac{S_{XY}}{S_{XX}} \right]$ **= 32718.18182/1432254.545 = 0.0228**

b) Estimated intercept, $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ **= 20.5454545 – (0.0228*603.6363636) = 6.756086989**

Excel Worksheet:



xbar=	ybar=	Sxy =	Sxx =	Syy =
603.6363636	20.5454545	32718.18182	1432254.545	1374.727273

Slope =	Intercept =
B1 hat =	B0 hat =
Sxy/Sxx =	ybar - b1 xbar
0.02284383	6.756086989

5. State the fitted line and interpret it.

$\hat{Y} = 6.76 + 0.02 * X$

A unit increase in X = per capita consumption of cigarettes (in 1930) is estimated to be associated with a .02 increase in Y = the number of lung cancer cases per 100,000 in 1950.

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 $= \sum_{i=1}^{11} \left(Y_i - \overline{Y}\right)^2$ **= 1374.727273**

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

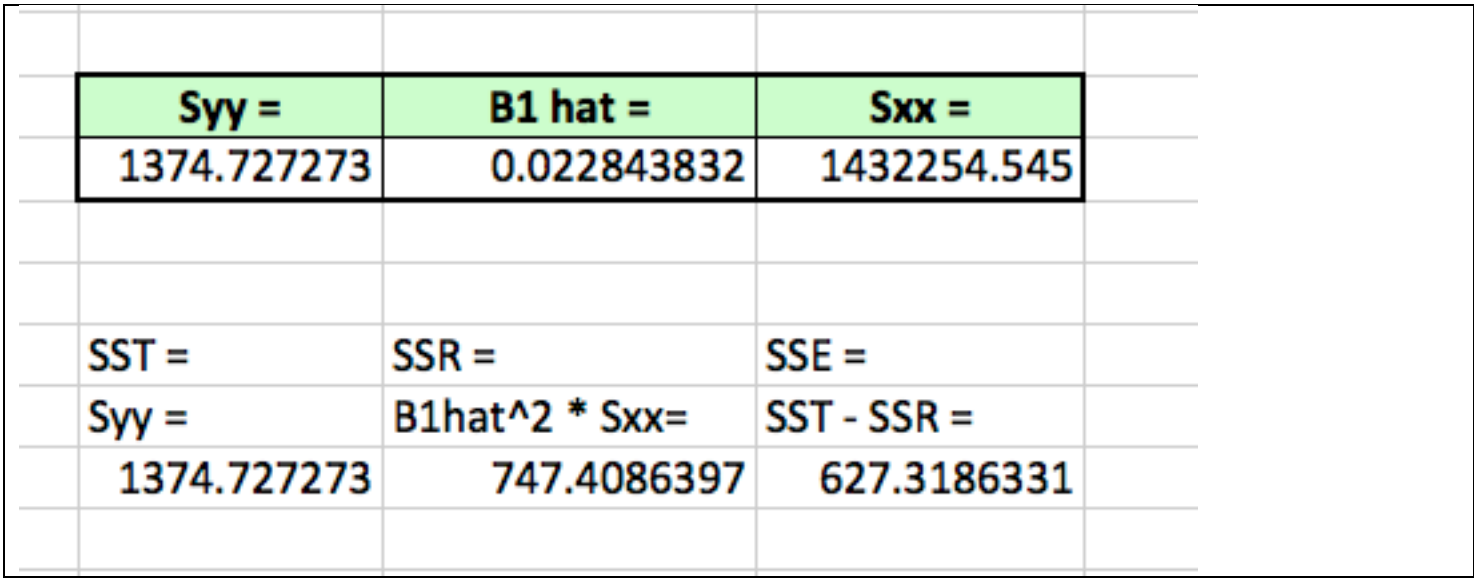
b) Regression sum of squares = SSR = $\sum_{i=1}^{11} \left(\hat{Y}_i - \overline{Y}\right)^2 = \hat{\beta}_1^2 \sum_{i=1}^{11} \left(X_i - \overline{X}\right)^2$ **= 747.4086397**

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

c) Error sum of squares = SSE = $\sum_{i=1}^{11} \left(Y_i - \hat{Y}\right)^2 = SST - SSR$ **= 627.3186331**

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

Excel Worksheet:



Syy =	B1 hat =	Sxx =
1374.727273	0.022843832	1432254.545

SST =	SSR =	SSE =
Syy =	B1hat^2 * Sxx=	SST - SSR =
1374.727273	747.4086397	627.3186331

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

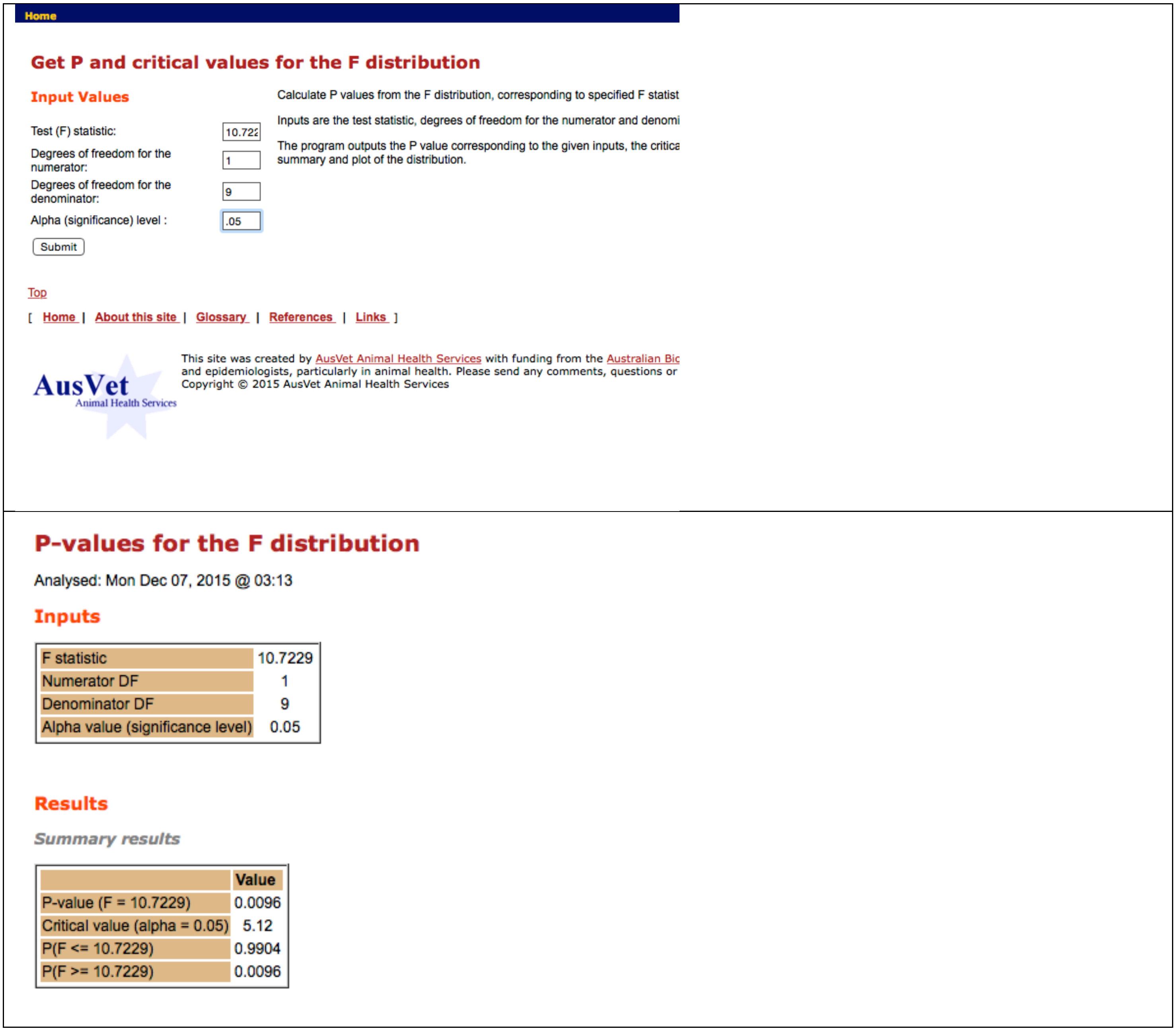
Source	df	Sum of Squares	Mean Square	F-Statistic
Regression	1	$SSR = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$ = 747.4086	SSR/1 = 747.4086	747.4086/ 69.70207 = 10.723
Error	(n-2) = 9	$SSE = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$ = 627.3186	SSE/(n-2) = 69.70207	
Total, corrected	(n-1) = 10	$SST = \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$ = 1374.7273		

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

8. Perform and interpret the overall F test.

Note – I used the Epi-Tools applet. You may have used a different one. As we saw previously, this particular calculator for the F-distribution (while quite thorough) requires that you input an alpha value, even if this is not of interest to you.

The F-test of the overall regression tests the null hypothesis that the slope is zero. Assumption of the null hypothesis model to these data yielded an observed F-statistic value of 10.723 with degrees of freedom 1 and 9. The achieved significance level (p-value) is .0096, representing a very unlikely result. The null hypothesis is rejected. Conclude that the fitted straight line model explains statistically significantly more of the variability in lung cancer cases than the model defined by the average.



Get P and critical values for the F distribution

Input Values

Calculate P values from the F distribution, corresponding to specified F statist

Inputs are the test statistic, degrees of freedom for the numerator and denomi

The program outputs the P value corresponding to the given inputs, the critica summary and plot of the distribution.

Test (F) statistic:	10.722
Degrees of freedom for the numerator:	1
Degrees of freedom for the denominator:	9
Alpha (significance) level :	.05

This site was created by AusVet Animal Health Services with funding from the Australian Bic and epidemiologists, particularly in animal health. Please send any comments, questions or Copyright © 2015 AusVet Animal Health Services

P-values for the F distribution

Analysed: Mon Dec 07, 2015 @ 03:13

Inputs

F statistic	10.7229
Numerator DF	1
Denominator DF	9
Alpha value (significance level)	0.05

Results

Summary results

	Value
P-value (F = 10.7229)	0.0096
Critical value (alpha = 0.05)	5.12
P(F <= 10.7229)	0.9904
P(F >= 10.7229)	0.0096

http://epitools.ausvet.com.au/content.php?page=f_dist