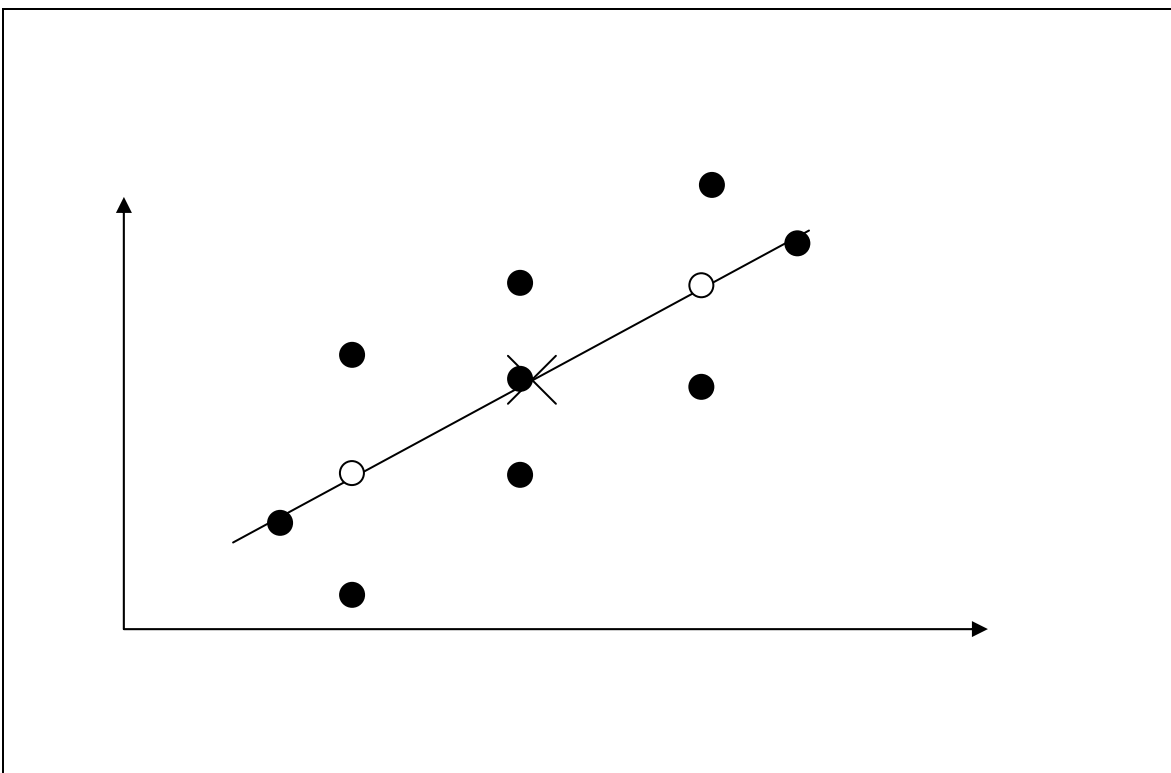


**Unit 2 – Regression and Correlation**  
**Practice Problems**

**Due: February 13, 2012**

1. A regression analysis of measurements of a dependent variable Y on an independent variable X produces a statistically significant association between X and Y. Drawing upon your understanding of biostatistics, epidemiology, and the scientific method, etc – see how many explanations you can offer for this finding. *Hint – I get seven (7)*
2. Below is a figure summarizing some data for which a simple linear regression analysis has been performed. The point denoted X that appears on the line is (x,y). The two points indicated by open circles were NOT included in the original analysis.



Multiple Choice (Choose ONE):

Suppose the two points represented by the open circles are added to the data set and the regression analysis is repeated. What is the effect of adding these points on:

1. The estimated slope.  
 (a) increase  
 (b) decrease  
 (c) no change
  
2. The residual sum of squares.  
 (a) increase  
 (b) decrease  
 (c) no change
  
3. The degrees of freedom.  
 (a) increase  
 (b) decrease  
 (c) no change
  
4. Standard error of the estimated slope.  
 (a) increase  
 (b) decrease  
 (c) no change
  
5. The predicted value of  $y$  at  $x=24$ .  
 (a) increase  
 (b) decrease  
 (c) no change

3. . The course website page [REGRESSION AND CORRELATION](#) will have some examples of code to produce regression analyses in STATA and SAS.

The data in the table below are values of boiling points (Y) and temperature (X). Carry out an exploratory analysis to determine whether the relationship between temperature and boiling point is better represented using

- (i)  $Y = \beta_0 + \beta_1 X$  or  
 (ii)  $100 \log_{10}(Y) = \beta'_0 + \beta'_1 X$

In developing your answer, use whatever statistical software you like. Try your hand at producing

- (a) Estimates of the regression line parameters  
 (b) Analysis of variance tables  
 (c)  $R^2$   
 (d) Scatter plot with overlay of fitted line.

Complete your answer with a one paragraph text that is an interpretation of your work. Take your time with this and have fun.

X=Temp	Y=Boiling Pt	X=Temp	Y=Boiling Pt	X=Temp	Y=Boiling Pt
210.8	29.211	193.6	20.212	184.1	16.817
210.1	28.559	191.4	19.758	183.2	16.385
208.4	27.972	191.1	19.490	182.4	16.235
202.5	24.697	190.6	19.386	181.9	16.106
200.6	23.726	189.5	18.869	181.9	15.928
200.1	23.369	188.8	18.356	181.0	15.919
199.5	20.030	188.5	18.507	180.6	15.376
197.0	21.892	185.7	12.267		
196.4	21.928	186.0	17.221		
196.3	21.654	185.6	17.062		
195.6	21.605	184.1	16.959		
193.4	20.480	184.6	16.881		

4. A psychiatrist wants to know whether the level of pathology (Y) in psychotic patients 6 months after treatment could be predicted with reasonable accuracy from knowledge of pretreatment symptom ratings of thinking disturbance ( $X_1$ ) and hostile suspiciousness ( $X_2$ ).

(a) The least squares estimation equation involving both independent variables is given by

$$Y = -0.628 + 23.639(X_1) - 7.147(X_2)$$

Using this equation, determine the predicted level of pathology (Y) for a patient with pretreatment scores of 2.80 on thinking disturbance and 7.0 on hostile suspiciousness. How does the predicted value obtained compare with the actual value of 25 observed for this patient?

(b) Using the analysis of variance tables below, carry out the overall regression F tests for models containing both  $X_1$  and  $X_2$ ,  $X_1$  alone, and  $X_2$  alone.

Source	DF	SS
Regression on $X_1$	1	1546
Residual	51	12246

Source	DF	SS
Regression on $X_2$	1	160
Residual	51	13632

Source	DF	SS
Regression on $X_1, X_2$	2	2784
Residual	50	11008

- (c) Based on your results in part (b), how would you rate the importance of the two variables in predicting Y?
- (d) What are the  $R^2$  values for the three regressions referred to in part (b)?
- (e) What is the best model involving either one or both of the two independent variables?

5. In an experiment to describe the toxic action of a certain chemical on silkworm larvae, the relationship of  $\log_{10}(\text{dose})$  and  $\log_{10}(\text{larva weight})$  to  $\log_{10}(\text{survival})$  was sought. The data, obtained by feeding each larva a precisely measured dose of the chemical in an aqueous solution and then recording the survival time (ie time until death) are given in the table. Also given are relevant computer results and the analysis of variance table.

Larva	1	2	3	4	5	6	7	8
$Y = \log_{10}(\text{survival time})$	2.836	2.966	2.687	2.679	2.827	2.442	2.421	2.602
$X_1 = \log_{10}(\text{dose})$	0.150	0.214	0.487	0.509	0.570	0.593	0.640	0.781
$X_2 = \log_{10}(\text{weight})$	0.425	0.439	0.301	0.325	0.371	0.093	0.140	0.406

Larva	9	10	11	12	13	14	15
$Y = \log_{10}(\text{survival time})$	2.556	2.441	2.420	2.439	2.385	2.452	2.351
$X_1 = \log_{10}(\text{dose})$	0.739	0.832	0.865	0.904	0.942	1.090	1.194
$X_2 = \log_{10}(\text{weight})$	0.364	0.156	0.247	0.278	0.141	0.289	0.193

$$Y = 2.952 - 0.550 (X_1)$$

$$Y = 2.187 + 1.370 (X_2)$$

$$Y = 2.593 - 0.381 (X_1) + .871 (X_2)$$

Source	DF	SS
Regression on $X_1$	1	0.3633
Residual	13	0.1480

Source	DF	SS
Regression on $X_2$	1	0.3367
Residual	13	0.1746

Source	DF	SS
Regression on $X_1, X_2$	2	0.4642
Residual	12	0.0471

- (a) Test for the significance of the overall regression involving both independent variables  $X_1$  and  $X_2$ .
- (b) Test to see whether using  $X_1$  alone significantly helps in predicting survival time.
- (c) Test to see whether using  $X_2$  alone significantly helps in predicting survival time.
- (d) Compute  $R^2$  for each of the three models.
- (e) Which independent predictor do you consider to be the best single predictor of survival time?
- (f) Which model involving one or both of the independent predictors do you prefer and why?

6. Using whatever software package you like, try your hand at reproducing the analysis of variance tables you worked with in problem #5.

7. An educator examined the relationship between number of hours devoted to reading each week ( $Y$ ) and the independent variables social class ( $X_1$ ), number of years school completed ( $X_2$ ), and reading speed measured by pages read per hour ( $X_3$ ). The analysis of variance table obtained from a stepwise regression analysis on data for a sample of 19 women over the age of 60 is shown.

Source		DF	SS
Regression	$(X_3)$	1	1058.628
	$(X_2 X_3)$	1	183.743
	$(X_1 X_2, X_3)$	1	37.982
Residual		15	363.300

- (a) Test the significance of each variable as it enters the model.
- (b) Test  $H_0: \beta_1 = \beta_2 = 0$  in the model  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + E$ .
- (c) Why can't we test  $H_0: \beta_1 = \beta_3 = 0$  using the ANOVA table given? What formula would you use for this test?
- (d) What is your overall evaluation concerning the appropriate model to use given the results in parts (a) and (b)?

8. Consider the following analysis of variance table.

Source		DF	SSQ
Regression	( $X_1$ )	1	18,953.04
	( $X_3 X_1$ )	1	7,010.03
	( $X_2 X_1, X_3$ )	1	10.93
Residual		16	2,248.23
			28,222.23

Using a type I error of 0.05,

(a) Provide a test to compare the following two models:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + E. \text{ VERSUS}$$

$$Y = \beta_0 + \beta_1 X_1 + E.$$

(b) Provide a test to compare the following two models:

$$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + E. \text{ VERSUS}$$

$$Y = \beta_0 + E.$$

(c) State which two models are being compared in computing:

$$F = \frac{(18,953.04 + 7,010.03 + 10.93)/3}{(2,248.23)/16}$$