Laboratory 6 – Bivariate Data: Linear Least-Squares Regression
Assigned Week of November 11, 2013
Due Week of November 18, 2013

I — Introduction:

In Laboratory 5, you dealt with a single variable and its measured distribution using univariate statistics. Engineering problems often involve relationships between two or more measured variables. In this laboratory we shall use bivariate statistics to test relationships between two coupled variables. One of the more useful tools to determine relationships between bivariate data is least squares regression. In this case, an equation is proposed as a possible relationship connecting the bivariate data, and this equation is tested using the method of least-squares regression. If the dependent variable depends linearly on the parameters in the equation (*i.e.,* the function is linear in the unknown parameters), then we can use linear least-squares regression. Linear least-squares regression is the focus of this laboratory exercise.

Regression is used to test hypotheses that are expressed as mathematical expressions. If the model is physically based then passing these tests gives some confidence in the physical insight used to generate the model; if the model does not pass the test, then we know our picture is wrong. For some applications, the model does not need to be sophisticated. As an example, we know from experiment that the strength of steel increases with carbon content. A complete understanding of why adding small quantities of carbon to steel increases its strength would require a lot of effort, but if all we want to do is to predict the strength of a new steel, then simple expressions determined from regressions to previous bivariate measurements might be all that is required.

II — Problem Statement:

Suppose that you are a Metallurgical Engineer and you have been asked to predict the strength of two new steels based on their carbon contents. You tabulate bivariate data on strength and carbon content (See Table 1). You find that the carbon content of one of the new steels (0.33 wt%) is within the range of previously measured steels, so you expect to be able to interpolate. For the other new steel, the carbon content is 1.0 wt%, which is greater than the values for the steels in the table, so you will have to extrapolate. In this laboratory exercise, you will predict the strengths of the new steels with simple functions, whose parameters will be determined by linear-least squares regressions.

Steels are alloys comprised of iron (Fe) and carbon (C), and are typically designated with a four-digit number by the *American Iron and Steel Institute* (AISI). The last two digits specify the nominal carbon content in these plain-carbon steels, as per the following example: AISI 1020 steel contains 0.20% carbon, where the 20 in 1020 becomes the 20 in 0.20%. The term 'plain carbon' means that carbon is an alloying element.

The data contained in Table 1 are examples of measured yield strengths for different steels. The yield strength for Steel Grade 1095 is not a measured value. For this exercise, the yield strength for Steel

Grade 1095 will be **500 plus the last two numbers of your student number**. So, if your student number is 100xxx123, then the yield strength you will use for Steel Grade 1095 is 523 MPa.

The data provided seems to suggest that a linear relationship might be a good first guess. Your task is to test this hypothesis and to determine a linear relationship between carbon content and measured strength for the steels found in Table 1. You will calculate the correlation coefficient and estimate the quality of your fitted result with the coefficient of determination.

Next you will test a third-order polynomial (cubic) and a sixth-order polynomial to see if these represent the data better than the linear relationship. You will calculate the correlation coefficient and estimate the quality of your fitted results with the coefficient of determination.

Finally, you will predict the strengths of the new steels using each of your determined equations. In your report you will compare the quality of the fitted equations and report the best predictions.

Table 1: Carbon Content and Measured Yield Strengths of Selected Plain Carbon Steels (Annealed Heat-Treated Condition)

Steel Grade	Carbon Content (Weight %)	Strength (MPa)
1015	0.15	315
1020	0.20	330
1022	0.22	358
1030	0.30	345
1040	0.40	415
1050	0.50	507
1060	0.60	483
1080	0.80	585
1095	0.95	**500 + last two digits of student #**

III — Steps and Calculations:

1. Determine the best-fit linear relationship (equation) for the strength of the steels as a function of their carbon content for the steels listed in Table 1 using the formulas provided in class (*i.e.,* calculate the slope and intercept for the best-fit straight line to the data.). You can do these calculations by hand, but you should find it easier to insert the equations into an Excel spreadsheet. Complete Table 2 below for the sums and averages and include it as an appendix to your report (use Excel).
2. Calculate the correlation coefficient using the formula provided in class. Record your calculations in the appendix of your report, and insert your results in Table 2.
3. Calculate the Coefficient of Determination from the ratio of the sum of the squares from the regression (SSR) and the total sum of squares (TSS). Compare this value with the correlation

coefficient calculated in Step 2. Record your calculations in the appendix to your report, and insert your results in Table 2.

4. Compare the results of Steps 2 and 3 with the same values calculated with Excel's CORREL(xstart:xend,ystart:yend) and RSQ(xstart:xend,ystart:yend) commands, respectively.
5. Now, use Excel's data analysis capability to calculate the best-fit slope, intercept, and coefficient of determination: under the Data tab, go to Data Analysis and then go to Regression. **Include the regression "Summary Output" generated by Excel in an appendix of your report.** Compare the results in the summary output with the values you determined with the formulas in Steps 1-3 above.
6. In the "Summary Output" of Step 5, find and report the 95% confidence limits for the determined slope and intercept for the regression.

Make sure you change the strength for Steel Type 1095 to 500 + the last two digits of your student number.

Table 2: Linear Model Calculations

Steel Type	Carbon Content (weight %)	Strength (MPa)	$x-\overline{x}$	$(x-\overline{x})^2$	$y-\overline{y}$	$(y-\overline{y})^2$	$(x-\overline{x})(y-\overline{y})$	Predicted Strengths $\hat{y}=mx+b$	$(\hat{y}-\overline{y})^2$	$(y-\hat{y})^2$
1015	0.15	315								
1020	0.2	330								
1022	0.22	358								
1030	0.3	345								
1040	0.4	415								
1050	0.5	507								
1060	0.6	483								
1080	0.8	585								
1095	0.95	500+*								

*** This value is 500 + the last two digits of your student number.**

Number of data		Slope	
$\sum_{i=1}^{n}(x_i-\overline{x})^2$		Intercept	
$\sum_{i=1}^{n}(y_i-\overline{y})^2$		Correlation Coefficient	
$\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})$		Coefficient of Determination	
$\overline{x}$			
$\overline{y}$			
s_x			
s_y			
TSS			
SSR			
SSE			

7. Create an X-Y scatter plot with the Excel plotting facility to represent the data graphically. Include labels for the axes, as well as a title. **Your name should appear in the graph title.** On the plot include the 'linear' trend line, the linear equation and R^2 value generated by Excel.

8. Using the linear regression equation, predict the ŷ-value (predicted strength) for each given x-value (carbon content) in Table 1, *i.e.,* enter the x data into the determined regression expression and solve for ŷ. Record these predicted values in Table 3 along with the value for R^2.

9. Produce two new graphs of the data. In one plot use a third-order polynomial trend line. In the second plot use a sixth-order polynomial for the trend line. Include labels for the axes, as well as a title. **Your name should appear in the graph title.** On each plot include the trend line, the equation and R^2 value generated by Excel.

10. Using the two regression equations determined in Step 9, predict the ŷ-value (predicted strength) for each given x-value (carbon content) in Table 1, *i.e.,* enter the x data into each of the determined regression expressions and solve for ŷ. Record these predicted values in Table 3 along with the values for R^2. Note: when displaying/copying the regression equations generated by Excel, insure that rounding does not occur in the equation's coefficients as this can produce equations for very different lines.

11. Complete Table 3 by predicting with your three regression equations the strength of the steel when the carbon content is 0.33 wt% (interpolation) and 1.0 wt% (extrapolation).

Table 3: Summary calculation table

Carbon Content (%)	**Measured Strength (MPa)**	Predicted Strength **(Linear)**	Predicted Strength **(Polynomial Degree=3)**	Predicted Strength **(Polynomial Degree=6)**
0.15	**315**			
0.2	**330**			
0.22	**358**			
0.3	**345**			
0.4	**415**			
0.5	**507**			
0.6	**483**			
0.8	**585**			
0.95	**...***			
R^2 value for the regression				
Interpolation: 0.33	-			
Extrapolation: 1.00	-			

*** This value is 500 + the last two digits of your student number.**

IV — Report Requirements and Deliverables:

Generate a brief one-page report for this laboratory according to the format discussed in Laboratory 1. What was the objective of this laboratory exercise? How was the objective approached, *i.e.,* what was the method? What were your results and how do they relate to the objective? What is your conclusion - was your objective realized (what are the relationships that you determined?

- Which regression equation best describes the relationship between steel strength and carbon content? What values would you predict for the strengths of the new steels? Explain your choices.

Deliverables Summary   ***The lab assignment includes the following:***
1. Title page   **2. One-page report**   **3. Completed Table 2.**   **4. Excel regression "Summary Output" (Step 5): 95% confidence intervals for slope and intercept.**   **5. Three plots with your name in the Title: a linear model, a third-order polynomial model and a sixth-order polynomial model. Each of the three plots includes the data in Table 1, the fitted equation and its R^2 value.**   **6. Completed Table 3.**   **7. IMPORTANT: submit your electronic version of your assignment to the specific folder***   **NOTE: You can present all of the plots on one single page of your appendix.**

***File name: "Lab Session_Student number.docx" (e.g. "C3_100812345.docx" for C3 Lab session)**

V — Submission and Timing:

Your report is to be submitted to the Teaching Assistant within the first 30 minutes of your next laboratory period. **LATE SUBMISSIONS WILL NOT BE ACCEPTED.**

VI — Marking:

Laboratory submissions will be marked on a 10-point scale: 9-10 (excellent); 7-8 (good); 5-6 (marginal); less than 5 (poor). **Be sure that you are familiar with the University's policy on plagiarism and academic integrity. Your instructors are obligated to report all suspected violations to the Associate Dean's office for investigation (see also chapter 14 at** http://calendar.carleton.ca/undergrad/regulations/academicregulationsoftheuniversity/ **).**