

# NOTE R: LINEAR REGRESSION

## CENTERING INPUT DATA (EXCEL 2000)

Generate a centered data set:

- a. Build a worksheet with the data. Set column A as the Y data and columns B to  $\{m+1\}$ <sup>1</sup>, as the X data. Reserve row 1 for labels that identify each of the X variables. Rows 2 thru  $\{n+1\}$  then will contain the original data.
  - b. In column A, row  $\{n+3\}$  enter =AVERAGE(A2:A $\{n+1\}$ ) . Formula copy across the columns.
  - c. In column A row  $\{n+5\}$  put in the formula =A2-A $\{n+3\}$  and formula copy across the columns.
  - d. Formula copies the selected cells down to row  $\{2n+4\}$ .
  - e. The centered data will be from rows  $\{n+5\}$  to  $\{2n+5\}$ .
  - f. The range for the Y data will be A $\{n+5\}$ :A $\{2n+5\}$ .
  - g. The range for the X variables will be B $\{n+5\}$ : $\{m+1\}$  $\{2n+5\}$ .
2. Set the Regression output to begin to the left of the centered data.
  3. The Regression routine output sheet will be correct for all variables except the intercept. The new intercept can be calculated as follows. Assume that the output value of the (centered) regression intercept is in cell Qw. This value should be close to zero.
    - a. In cell A $\{2n+7\}$ . = - A $\{n+3\}$  + Qw ‘The Y mean’
    - b. In cell B $\{2n+7\}$ . =Q $\{w+1\}$  \* B $\{n+3\}$
    - c. In cell C $\{2n+7\}$ . =Q $\{w+2\}$  \* C $\{n+3\}$
    - d. Continue across all variable columns.
    - e. In a blank cell put in SUM(A $\{2n+7\}$ :  $\{m+1\}$  $\{2n+7\}$ ). This will be the correct intercept.

In all the cases I investigated, Excel produced a satisfactory solution after centering the data about the means of each variable. The multi-colinearity problem was not a problem here, since in all the difficult cases with high multi-colinearity (Fillip and Longley), Excel came up with correct solutions.

In most cases, centering the data about the mean (an approximate mean is OK) will improve accuracy of the result. However this should not be done when the Data Analysis Regression routine or when any of the LINEST functions are used on a through-the-

---

<sup>1</sup> The  $\{ \dots \}$  is a notation whose value has to be converted either to a row number or an alphabetic character corresponding to a column designation. n is the number of “points” or observations in the data set, and m is the number of variables in the data set.

origin selected model. With a through-the-origin model, the absolute X values are essential to getting correct values. Here the centered values will give the wrong output. Note J describes how to correct the Regression routine output from the original, uncentered data, to give correct correlations and variance table values.