EXPLORATIONS

Activity 11

Inference for Correlation and Regression

Both the introduction and Topic 11 in Activity 3 discussed fitting a straight line to bivariate data. Topic 52 below extends this discussion to test if a significant relationship exists between the two variables and then calculate the confidence and predictive intervals.

Topic 53 extends the above to more than one independent variable and explains how to use program **A2MULREG**, which also automates the procedure in Topic 52.

Topic 52—Simple Linear Regression and Correlation (Hypothesis Test and Confidence and Predictive Intervals)

A study was conducted to investigate if there was a relationship between the length of time a student studies outside of class each week and the final grade in a course. A simple random sample of ten students from the course was used and is given below.

Student	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Hrs. Studied (x) L_1:	3.5	6	7	3	4.5	7.5	4	6.5	5.5	5
Final Grade (y) L_2:	75	95	83	69	77	93	73	87	78	86

Put hours in list **L_1** and grades in **L_2**, and then continue with the following procedure.

Activity 11, Inference for Correlation and Regression (cont.)

1. Set up a **scatter** plot.

   Set up **Plot1** as in Topic 7, and press [ZOOM] **9:ZoomStat** [TRACE] for screen 1. The [◄] [►] keys can be used to highlight each point.

2. Test the null hypothesis H_o: $\beta = 0$ and H_o: $\rho = 0$.

   a. Press [STAT] **<TESTS>E:LinRegTTest** for screen 2.

   b. Paste **L1** and **L2** for the **Xlist** and **Ylist** and paste **Y1** for the **RegEQ** with [VARS] **<Y-VARS>1:Function 1:Y1**.

      Note the alternate hypothesis is set at $\beta \neq 0$ and $\rho \neq 0$.

   c. Highlight **Calculate** at the bottom of screen 2 and press [ENTER] for the first screens 3 and 4, where:

      regression line = **y = a + bx** = **56.90909** + **4.70303**x
      correlation coefficient = **r = 0.82491**
      coefficient of determination = **r2 = 0.68048**

With a p-value of **0.003309** and **b** and **r** positive, we conclude that the slope of the population regression line β is significantly different from zero and that there is significant positive correlation between hours studied and final grade.

3. Find 95 percent confidence interval for β.

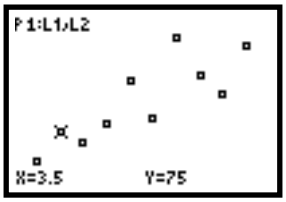
   t = (**b** − 0)/**S_b**; therefore, **S_b** = **b/t** = **4.70303/4.12766** = **1.1394**, as shown in screen 5. Notice you could paste **b** and **t** with [VARS] **5:Statistics.<EQ> 3:b** and [VARS] **5:Statistics.<TEST> 3:t**.

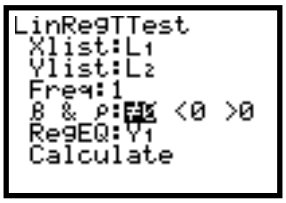
   a. Find the critical t-value from a table or by using the equation solver to solve **tcdf(X, E99, 8) = 0.025** for **X**, as explained at the end of Topic 34. (Degrees of freedom = n − 2 = 10 − 2 = 8.)

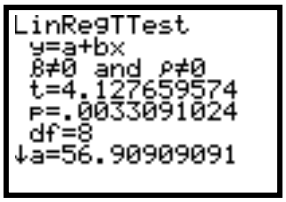
   b. To verify the critical value X = **2.306**, press [2nd] [DISTR] **5:tcdf(2.306** [,] **E99** [,] **8** [)] [ENTER] for **0.025** (see the last two lines in screen 5).

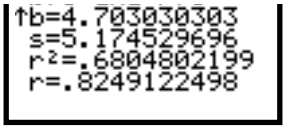
      The margin of error is **t** ∗ **S_b** = **2.627** = **E**, as shown in the first lines of screen 6.

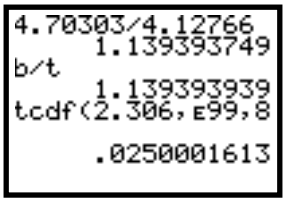
We are 95 percent confident that the slope of the population regression line is between 2.08 and 7.33. For each additional hour that a student studies, we expect the grade to increase from between 2.08 to 7.33 percentage points.

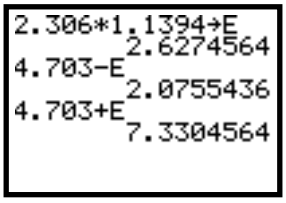
(1)


(2)


(3)


(4)


(5)


(6)


4. Plot data and regression line with point estimates when $X = 3.5$.

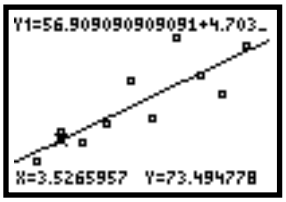
   a. With **Plot1** still set up from step 1 and with the regression equation automatically stored in **Y1** from step 2, press [TRACE] [▲] for the graph in screen 7. The cursor is flashing on the regression line.

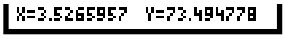
   b. Press [◀] until you are close to **3.5** (**3.5266** is the closest pixel (screen 8)).

   c. Type **3.5** and the large **X=3.5** appears at the bottom of the screen (screen 9).

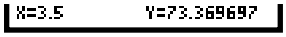
      Press [ENTER] for **Y= 73.369697** (screen 10).

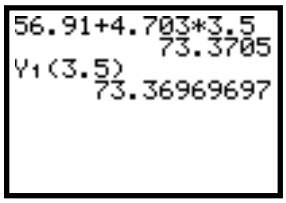
      You also could have entered **3.5** into the regression equation, as shown in screen 11.

(7)


(8)


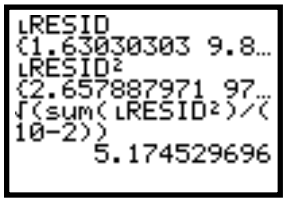
(9)


(10)


(11)


5. Calculate residuals and residual plots.

   Because all the points do not fall on the regression line, an interval estimate makes more sense than the point estimate used in step 4. A measure of the difference between the actual y-value of the data and the y-value on the regression curve for the same **x** is called the *residual*. (For the first point, **x = 3.5** and **y =75**, the regression line gives **Y1(3.5) = 73.3697**, so the residual is **75 – 73.3697 = 1.6303**.) The residuals for all the data points are automatically stored in list **LRESID** in step 2 (see the first two lines in screen 12).

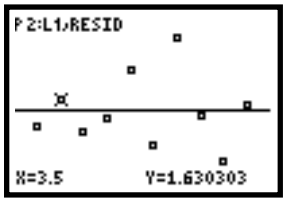
(12)


   A measure of the scatter of the points about the regression line is the square root of the sum of the residual squared divided by (**n – 2**), or **s = 5.1745**, as shown in the last line of screen 12 and in screen 4.

   a. Set up **Plot2** for a scatter plot with the Xlist = **L1** and Ylist = **LRESID** and with all other stat plots and Y= plots turned off.

   b. Press [ZOOM] **9**:ZoomStat [TRACE] for the plot in screen 13.

      Notice a fairly random pattern, but the residuals seem to get larger for longer study times.

(13)


6. Find 95 percent predictive and confidence intervals.

   a. For **X = 3.5** and the critical *t* value, **T = 2.306** (as calculated in step 3), you can calculate the predictive interval, as shown in screens 14 and 15 (other **X**s or interval levels could be used):

      s from [VARS] **5:Statistics <TEST> 0:s**.

      n and $\bar{x}$ from [VARS] **5:Statistics <XY> 1:n** and **2:** $\bar{x}$.

      Σx^2 and Σx from [VARS] **5:Statistics <Σ> 2:**Σx^2 and **1:**Σx.

   We are 95 percent confident (based on this small sample) that the grade obtained by a student who studies 3.5 hours a week is between about 60 and 87 percent.

   b. To calculate the confidence interval, use the [2nd] [ENTRY] feature to recall the lines, as in screen 16, and then delete the **1+** under the square root sign.

      As shown in screen 17, we obtain approximately 67 to 79 percent.

   The confidence interval is narrower than the predictive interval because this is the mean time we would predict for all students who study 3.5 hours (thus, by the *Central Limit Theorem*, the highs and the lows average out).

   Topic 53 automates this process.

(14)
```
3.5→X:2.306→T:Y₁
(X)→Y
          73.36969697
T*s√(1+1/n+n(X-x̄
)²/(n*Σx²-(Σx)²)
)→E
          13.33281413
```

(15)
```
Y-E
          60.03688284
Y+E
          86.7025111
```

(16)
```
3.5→X:2.306→T:Y₁
(X)→Y
          73.36969697
T*s√(1/n+n(X-x̄)²
/(n*Σx²-(Σx)²))→
E
          5.94812577
```

(17)
```
Y-E
          67.4215712
Y+E
          79.31782274
```

Topic 53—Multiple Regression and Program A2MULREG

To possibly improve the prediction capability of the regression equation developed in Topic 52 (which we will assume you are familiar), the age of the student (perhaps, related to motivation) will also be considered (see below).

	C_1	C_2	C_3
Student	**Y** (Grade)	**X1** (Study Hrs.)	**X2** (Age Yrs.)
1	75	3.5	20
2	95	6	19
3	83	7	36
4	69	3	21
5	77	4.5	27
6	93	7.5	24
7	73	4	22
8	87	6.5	34
9	78	5.5	23
10	86	5	25

Store the above data into a 10x3 matrix [D] as discussed in Topic 48 and partially shown in screen 18. The Y-values must be in column 1 (**C1**) of the matrix.

1. Set up program **A2MULREG.**

   Program **A2MULREG** is available from Texas Instruments over the internet (www.ti.com) or on disk (1-800-TI-CARES) and can be stored in your TI-83 with TI-GRAPH LINK. (The program listing is provided in Appendix B.)

   a. Press [PRGM] <EXEC>, highlight program **A2MULREG**, and then press [ENTER] to paste the name to the screen, as shown in screen 19.

   b. Press [ENTER] for the next screen (screen 20), which reminds you to put the data in matrix [D] and informs you that matrices [A] to [F] will be used by the program. To eliminate the fear of losing data, you can use matrices [G], [H], [I], and [J] for saving data.

      Notice the pause indicator in the upper right corner of the screen waiting for input or, in this case, for you to press [ENTER].

(18)
```
MATRIX[D] 10×3
[ 75      3.5     20      ]
[ 95      6       19      ]
[ 83      7       36      ]
[ 69      3       21      ]
[ 77      4.5     27      ]
[ 93      7.5     24      ]
[ 73      4       22      ]
```

Note: *The first two columns above were stored in L2 and L1 in Topic 52 so they could be transferred to matrix [D] (as discussed in Topic 18) using [2nd] [LIST] <OPS> **0**:Listmatr(L2 [,] **L1** [,] [D] [ENTER]. This gives a 10 x 2 matrix. Change this to 10 x 3, and enter the last column by using [▼].*

(19)
```
prgmA2MULREG
```

(20)
```
DATA IN MAT [D]
COL Y,X1,X2,..XN
Y MUST BE IN THE
1ST COL OF [D].

[A],[B],[C],[D],
[E]+[F] USED.
```

2. Make the correlation matrix.

   a. Press ENTER for the menu in screen 21, and select **2:CORR MATRIX** for screen 22.

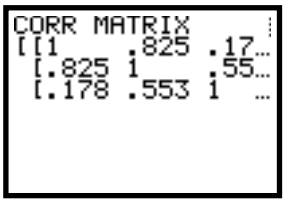
   b. View the rest of the matrix by pressing ▶.

   The simple linear correlation coefficient between **Y** and **X1** is **0.825** (as in Topic 52), between **Y** and **X2** is **0.178**, and between **X1** and **X2** is **0.553**.

   Again, notice the pause indicator. Pressing ENTER gives a **Done**.

(21)

(22)



3. Calculate simple linear regression. (**Y = B0 + B1 x 1**)

   To relate program **A2MULREG** to Topic 52, we will use only the first two columns of matrix [D]. The matrix could have been of order 10x2, but 10x3 is also acceptable because the last column is ignored for this step.

   a. Rerun program **A2MULREG**, and select **1:MULT REGRESSION** from the menu screen (screen 21) for screen 23.

   b. Enter **1** for **HOW MANY IND**ependent **VAR**iables, and then press ENTER.

   c. Enter **2** for **COL**umn of independent **VAR**iable. Remember **Y** is in column 1 and **X1** is in column 2.

   Because there is only one independent variable, you have the option of automatically plotting the scatter of points with the least square regression line, as shown in screen 26 and in Topic 52 (screens 1 and 7).

   d. Press ENTER**.**

   After a brief wait while the busy indicator is on in the upper right corner of the screen, the output in screen 27 appears, and the indicator changes to pause.

   p-value = **0.003**, r^2 = **0.6805** = **R-SQ**, **s** = **5.1745** and $\sqrt{\mathbf{F}} = \sqrt{(\mathbf{17.04})} = \mathbf{4.13} = \mathbf{t}$, all as in Topic 52. (Screens 3 and 4)

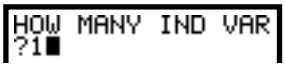
   F = (**456.193939**/**1**)/(**214.206060**/**8**)

   = (**456.193939**)/(**26.7757575**) = **17.04**

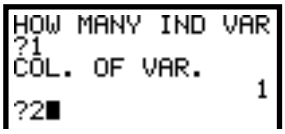
   with MSR = **456.193939** and MSE = **26.7757575**.

*Note: If no calculations have been done on the home screen since program **A2MULREG** was last run, pressing ENTER will restart the program.*

(23)

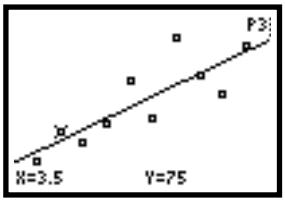


(24)

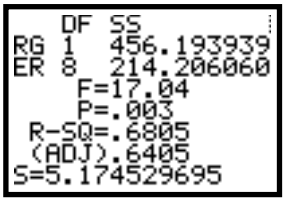


(25)

(26)



(27)



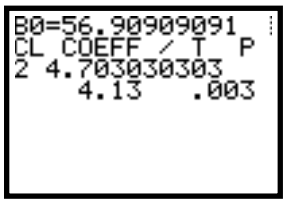
e. Press [ENTER] and the output is completed with **B0** = a = **56.9091**.

The **COEFF**icient of the **Co**L**u**mn 2 is **B1** = **b** = **4.70303**. Therefore, the regression equation is **Y** = **56.9091** + **4.70303**x, as in Topic 52 (screens 3 and 4). The **t** and **p** are given in the last line in screen 28.

The **t** of **4.13** is under the coefficient used to test the hypothesis $\beta 1 = 0$. The **p** value of **0.003** is beside the t-value it goes with.

In the simple linear regression case, the t-value and the F-value are directly related because there is only one independent variable. In the multiple regression case, there are multiple t -values and none are directly related to the F-value.

(28)



4. Find confidence and predictive intervals. (**Y = B0 + B1 x 1**)

a. After finding the simple linear regression, press [ENTER].

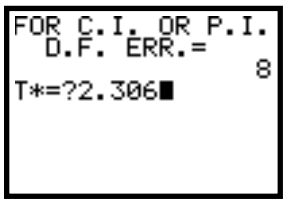
This reveals the MAIN MENU in screen 29 for the **Multiple Regression** option of program **A2MULREG**.

(29)

b. Select **1:CONF+PRI INTER** for input screen 30.

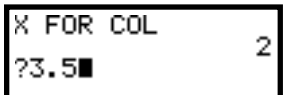
Enter **2.306** for the critical value for 95 percent intervals with 8 degrees of freedom (10 - 2 = 8), as in Topic 52 (screen 5).

(30)



c. Press [ENTER], and type **3.5** for the number of hours studied for which you want to predict the final grade earned (screen 31).

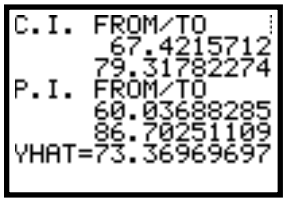
(31)



d. Press [ENTER] again to reveal the confidence interval, the predictive interval, and the point estimate **73.37** percent; all as in Topic 52 (screens 14-17), but this time, automated (screen 32).

Pressing [ENTER] again gives you the option of either entering another **X** or returning to the MAIN MENU.

(32)



5. Plot residuals. (**Y = B0 + B1 x 1**)

   a. From the **MAIN MENU**, select **2:RESIDUALS** for the menu in screen 33, which provides the option of plotting the residuals, plotting the standard residuals, or calculating the Durbin-Watson statistic.

   b. Select **1:RESIDUAL PLOT** for the next option (screen 34).

   c. Select **2 VS AN IND VAR** for the prompt "**WHAT COL?**". Enter **2** at this prompt for **X2**. The same residual plot appears as shown in screen 13 in Topic 52.

   d. Press ENTER and repeat the process for **1 VS YHAT** for the plot, as shown in screen 35.

      Notice the plot has the same scatter of points. Also, notice the Y-values are the same as the previous screen, but the X-values are now the result of entering **X1** into the regression equation (**YHAT**s) and not the **X1**s themselves.

6. View residual output. (**Y = B0 + B1 x 1**)

   If you select the **5:QUIT** option after pressing ENTER, screen 36 appears. It informs you where certain values can be observed.

   Press **STAT 1:Edit** for the first six lists, as shown in screens 37 and 38.

   If **2:RESIDUALS** is *not* selected from the MAIN MENU, then the values will *not* be listed as above. If **3:NEW MODEL** is selected, then even if the values had been calculated as above, they would now be cleared for the new model.

(33)
```
RES OR STAND RES
1:RESIDUAL PLOT
2:STAND.RES.PLOT
3:DURBIN WATSON
4:MAIN MENU
5:QUIT
```

(34)
```
PLOT OF RESID
1:VS YHAT.
2:VS AN IND VAR.
3:VS ROW NUMBER.
4:PREVIOUS MENU
5:MAIN MENU
6:QUIT
```

(35)
```
P3
X=73.369697  Y=1.630303
```

(36)
```
UNDER STAT EDIT
Y IN LYVAL,THEN
LYHAT, LRES, LSRES
LLEVER, LCOOKD
FOR LEVERAGE AND
COOK DISTANCE
                Done
```

(37)

YVAL	YHAT	RES 1
75	73.37	1.6303
95	85.127	9.8727
83	89.83	-6.83
69	71.018	-2.018
77	78.073	-1.073
93	92.182	.81818
73	75.721	-2.721

YVAL(1) =75

(38)

SRES	LEVER	COOKD 6
.36344	.24848	.04171
2.0423	.12727	.30415
-1.523	.24848	.38329
-.4821	.34545	.06133
-.2219	.12727	.00359
.19544	.34545	.01008
-.5792	.17576	.03577

COOKD(1)=.02183681...

7. Calculate multiple regression. (**Y = B0 + B1 x 1 + B2 x 2**)

   If you have selected **3:NEW MODEL** from the MAIN MENU or **1:MULT REGRESSION** after starting program **A2MULREG**, you will get a series of input screens like those condensed in screen 39.

   a. Select two independent variables with **X1** in column 2 and **X2** in column 3 of matrix [D].

   b. Press ENTER for screen 40, which shows very significant overall regression with a **p** value of **0.004**. **R-SQ** has been increased to **0.7917** from **0.6805** with only **X1** in the model and **s** decreased to **4.466** from **5.1745**.

   c. Press ENTER again for screen 41 with the regression equation **Y = 65.3829** + **5.9641X1** – **0.6014X2**.

   Testing H0: $\beta 1 = 0$ against $\beta 1 \neq 0$ brings a **t** = **5.05** with a **p** value = **0.001**.

   Testing H0: $\beta 2 = 0$ against $\beta 2 \neq 0$ brings a **t** = **–1.93** with a **p** value = **0.094**.

(39)
```
HOW MANY IND VAR
?2
COL. OF VAR.
                1
?2
COL. OF VAR.
                2
?3
```

(40)
```
   DF SS
RG 2  530.767417
ER 7  139.632582
    F=13.3
    P=.004
 R-SQ=.7917
 (ADJ).7322
S=4.466263744
```

(41)
```
B0=65.38293558
CL COEFF / T  P
2 5.964084176
    5.05   .001
3 -.6013696214
   -1.93   .094
```

Note: *Age (X2) does not add significantly to the model (at a = 0.05 < 0.094).*

8. Find confidence and predictive intervals. (**Y = B0 + B1 x 1 + B2 x 2**)

   a. After completing step 7 for multiple regression, press ENTER to reveal the MAIN MENU for the **Multiple Regression** option of program **A2MULREG** (see screen 42).

   b. Select **1:CONF+PRI INTER** for the input shown in screen 43.

      Enter **2.365** for the critical value for 95 percent intervals with degrees of freedom = 10 - 3 = 7 from a table or as done in Topic 35 with the equation solver.

(42)
```
MAIN MENU
1:CONF+PRI INTER
2:RESIDUALS
3:NEW MODEL
4:QUIT
```

(43)
```
FOR C.I. OR P.I.
  D.F. ERR.=
                7
T*=?2.365
```

c. Press ENTER, and then type **3.5** for **X1** (in **COL 2**), the number of hours studied weekly by the student of interest (screen 44).

d. Press ENTER, and then type **25** for **X2** (in **COL 3**), the age of the student of interest (screen 44).

e. Press ENTER again to reveal the confidence interval, the predictive interval, and the point estimate of **71** percent (screen 45) compared to the point estimate of 73 percent without age in the model. The interval widths also decreased a bit.

Pressing ENTER again now gives you the option of entering another **X** or returning to the MAIN MENU (screen 42).

(44)
```
X FOR COL
                2
?3.5
X FOR COL
                3
?25
```

(45)
```
C.I. FROM/TO
     65.33925489
     77.10672442
P.I. FROM/TO
     59.13211472
     83.31386459
YHAT=71.22298965
```

9. Plot residuals.

a. From the MAIN MENU, press **2:RESIDUALS** for the menu in screen 33.

b. Select **1:RESIDUAL PLOT** for the options in screen 34, and select **1 VS YHAT** for the plot shown in screen 46.

If you now press 2nd [QUIT], the residual output is entered in the stat editor, as shown in screens 47 and 48.

(46)
```
                              P3
X=74.229838  Y=.77016224
```

(47)
```
YVAL  | YHAT   | RES    1
95    | 74.23  | .77016
95    | 89.741 | 5.2586
83    | 85.482 | -2.482
69    | 70.646 | -1.646
77    | 75.984 | 1.0157
93    | 95.681 | -2.681
73    | 76.009 | -3.009
YVAL(1) =75
```

(48)
```
SRES    | LEVER  | COOKD  6
.20024  | .25841 | 0.00465
1.5365  | .41277 | .55311
-.7876  | .502   | .20841
-.4563  | .34731 | .03693
.25202  | .18576 | .00483
-.8571  | .50962 | .25448
-.7426  | .17687 | .0395
COOKD(1)=.00465719...
```

Limitations of A2MULREG

Program **A2MULREG** can handle many variables and data points sufficient for most Introductory Statistics text data sets, but is limited by the memory of the TI-83. For large data sets, you might want to clear some items saved in memory.

Remember, the columns of a matrix and a list can be interchanged (as in Topic 18), making data transformations possible.