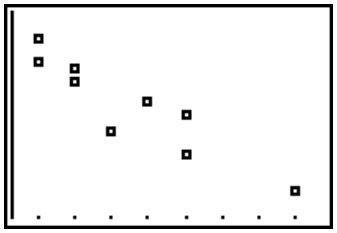
Chapter 15 – Multiple Regression

Multiple regression is an extension of the linear regression already studied where we create a model to explain a response variable based on more than one predictor. Just as with linear regression, we will want to examine how well the predictors determine the response, individually and as a group, by testing the utility of the model and create confidence intervals for slopes, mean response, and predictions of new responses. TI-83/84 calculators do not have a built-in multiple regression capability. TI-89 calculators do have this ability. If you are using a TI-83 or -84, the author of this manual has written a program (provided on the text's web site with a listing at the end of this chapter) called MULREG to do this.

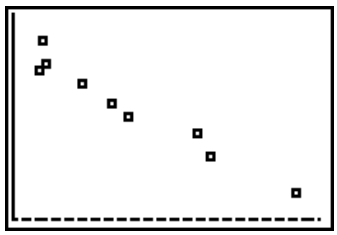
How well do age and mileage determine the value of a used Corvette? The author chose a random sample of ten used Corvettes advertised on autos.msn.com. The data are below.

Age (Years)	Miles (1000s)	Price ($1000)
3	46	27
1	11	43
2	20	35.5
1	11.5	39
8	69	16.5
5	49	23
2	10	38
4	27	32
5	30.5	30
3	46	27

We first examine plots of each predictor variable against Price. The plot against age is linear, and decreasing as expected (we expect older cars to cost less). The regression equation for this relationship is *Carprice* = 42.44 – 3.33**Age*, with r^2 = 80.6%. This suggests the average price of a used Corvette goes down $3330 each year.

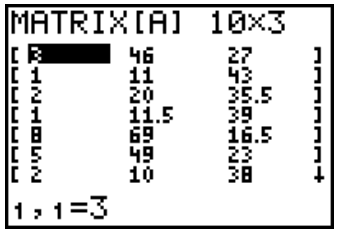


The plot of price against mileage is also linear and decreasing with even less scatter than in the other plot. The regression equation for this relationship was found to be *Carprice* = 43.77 – 0.40**Miles*, with r^2 = 95.5%. This suggests the average price of a used Corvette will decrease $400 for every 1000 miles it has been driven.

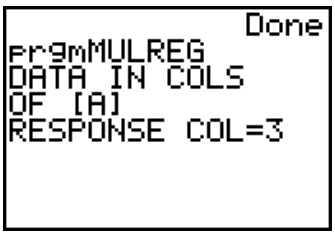


TI-83/84 Procedure

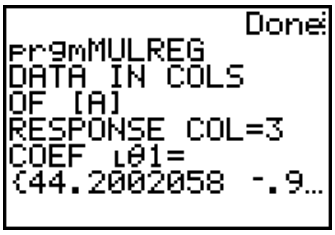
First the data are entered into Matrix [A]. On the TI-83, press MATRX; if you have a TI-83+ or TI-84, Matrix is 2nd x^{-1}. Arrow to EDIT and press ENTER to select matrix [A]. Enter the number of observations (or rows; here that is 10) and the number of columns (3). Type in the data, pressing ENTER after each number, across the rows of the matrix. Press 2nd MODE (Quit) to exit the editor.



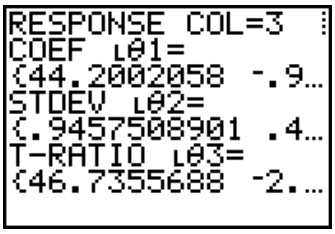
With the program transferred to the calculator, pressing PRGM will give the list of all programs stored. Select MULREG, then press ENTER to start it running. You will first be asked which column of the matrix has the response (Y) variable. In our example, price is in the third column. Press ENTER after the response.



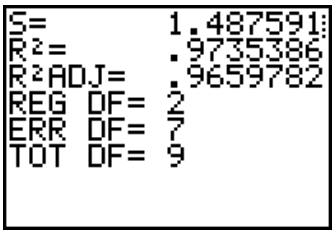
Here is the first portion of the output, the coefficients. These are displayed with the program paused so you can use the right arrow to scroll through them. They are also stored in list θ1 which can be accessed under the LIST menu (2nd STAT). Press ENTER to resume execution of the program. We find the equation of the model is $Price = 44.200 - 0.943 * Age - 0.309 * KMiles$.



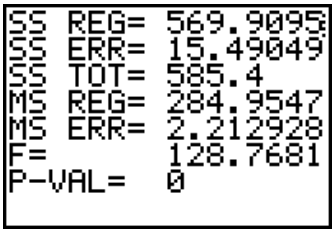
The next quantities displayed are the standard deviations of the coefficient estimates. As with the coefficients themselves, the calculator is paused to allow scrolling through the list. These are stored in list θ2 for further use. After pressing ENTER, the *t* statistics for each coefficient are displayed in a like manner, followed by the p-values for testing the hypotheses H_0: $\beta_i = 0$ against H_A: $\beta_i \neq 0$.



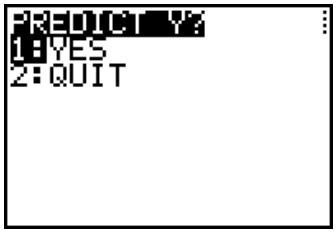
The next screen first displays S, the standard deviation of the residuals, then R^2, the coefficient of determination, 97.4%, which is how much of the variation in response (price) is explained by the model (in this case the age of the car and its mileage). The next quantity shown is the adjusted R^2. Since R^2 can never decrease when additional variables are added into a model, this quantity is "penalized" for additional variables which do not significantly help explain variation in Y, so it will go down if this is the case. Adjusted R^2 is always less than the regular R^2. We also see the degrees of freedom which are associated with the F-test for overall utility of the model. Press ENTER to resume execution of the program.



This screen shows the sums of squares and the F statistic for the overall significance (utility) of the model, along with its p-value. Here the p-value (to 5 decimal places) is 0 indicates there is a significant relationship between price and its two predictors.

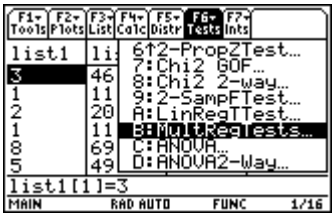


Now we are asked if we want to use the model to predict a value based on the equation, or quit. Make the appropriate choice.

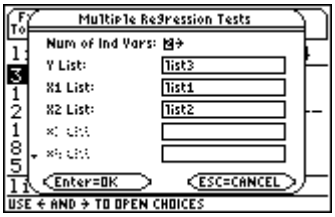


TI-89 Procedure

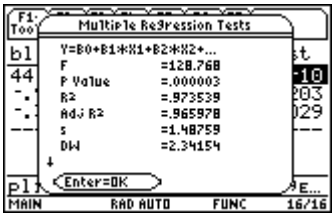
From the Statistics list editor, press 2nd F1 (Tests). Select menu option B:MultRegTests.



The input screen first asks how many independent variables there are. Use the right arrow to change this to the proper number. Enter the list names for the *Y* list and the *X* lists using 2nd [-] (VAR-LINK). Press [ENTER] to perform the calculations when all list names have been entered.



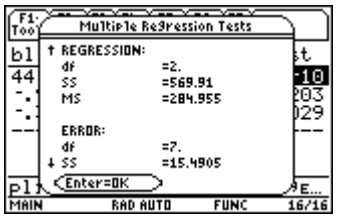
Here is the first portion of the output. The first line is the form of the regression equation. The second and third lines give the F statistic for the overall significance (utility) of the model, along with its p-value. Here the p-value of 0.000003 indicates there is a significant relationship between price and its two predictors. R^2 which is 97.3 % in this situation is the amount of variation in price which is explained by the two predictors (age and mileage).



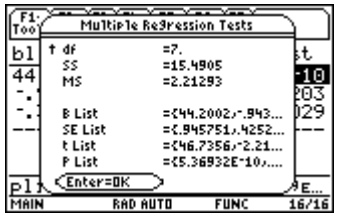
The next quantity is the adjusted R^2. Since R^2 can never decrease when additional variables are added into a model, this quantity is "penalized" for additional variables which do not significantly help explain variation in Y, so it will go down if this is the case. Adjusted R^2 is always less than the regular R^2. *S* is the standard deviation of the residuals.

DW is the value of the Durbin-Watson statistic which measures the amount of correlation in the residuals and is useful for data which are time series (data that have been collected through time). If the residuals are uncorrelated, this statistics will be about 2 (as it is here); if there is strong positive correlation in the residuals, DW will be close to 0; if the correlation is strongly negative, DW will be close to 4. Since these data are not a time series, DW is meaningless for our example.

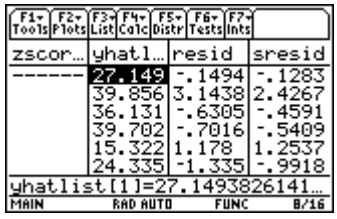
Pressing the down arrow we find the components for regression and error which are used in computing the F statistic. The F statistic for regression is the MS(Reg)/MS(Error) where the Regression Mean square functions just like the treatment (factor) mean square in ANOVA.



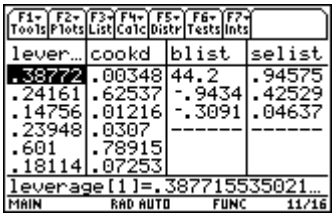
Finally we see some of the entries in new lists that have been created. The complete lists will be seen when [ENTER] is pressed. `Blist` contains the estimated intercept and coefficients; `SE list` is the list of standard errors for the coefficients which can be used to create confidence intervals for true slopes; `t list` gives values of the *t*-statistics for hypothesis tests about the slopes and intercept; `P list` gives the p-values for the tests of the hypotheses H_0: $\beta_i = 0$ against H_A: $\beta_i \neq 0$. If the assumed alternate is 1-tailed, divide these p-values by 2 to get the appropriate p-value for your test.



After pressing [ENTER] we see several new lists that have been added into the editor. `Yhatlist` is the list of predicted values for each observation in the dataset based on the model ($yhat_i = b_0 + b_1 x_{1i} + b_2 x_{2i}$ in this model); `resid` is the list of residuals $e_i = y_i - yhat_i$. `Sresid` is a list of standardized residuals obtained by dividing each one by *S*, since they have mean 0. If the normal model assumption for the residuals is valid, these will be N(0, 1).

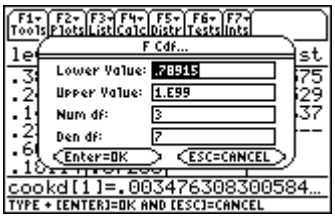


Pressing the right arrow we find yet more lists. Leverage is a measure of how influential the data point is. These values range from 0 to 1. The closer to 1, the more influential (more of an outlier in its x values) the point is in determining the slope and intercept of the fitted equation. Values greater than $2p/n$ where n is the number of data points and p is the number of parameters in the model are considered highly influential. Here, $n = 10$ and $p = 3$, so any value grater than 0.6 will designate an observation as highly influential. This indicates in our example that the point for the 8-year-old car is influential.

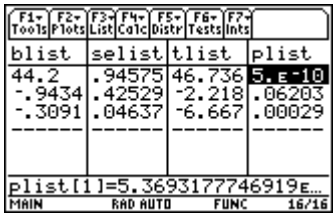


Cook's Distance in the next column is another measure of the influence of a data point in terms of both its x and y values. Its value depends on both the size of the residual and the leverage. The i^{th} case can be influential if it has a large residual and only moderate leverage, or has a large leverage value and a moderate residual, or both large residual and leverage.

To assess the relative magnitude of these values, one can compare them against critical values of an F distribution with p and $n - p$ degrees of freedom or use menu selection `A: F Cdf` from the F5 (`Distr`) menu. The largest value in the list is for the Corvette data is again for the fifth observation (the 8-year-old). This is the input screen. We are finding the area above 0.78915 which is the largest value in the list. The result is 0.5372, which indicates this is not unusual, so this point is not influential.



After pressing the right arrow still more, we find the last of the output lists. `Blist` is the list of coefficients. We finally see the fitted regression equation: $Price = 44.2 - 0.94 * Age - 0.31 * Miles$. We interpret the coefficients in the following manner: price declines \$940 on average for each year of age when mileage is the same; for cars of a given age, every additional 1000 miles reduces average price \$310. The coefficient of miles is similar to that obtained from the simple regression (\$400) but the decrease for age is much less than the value for the simple regression (\$3300).

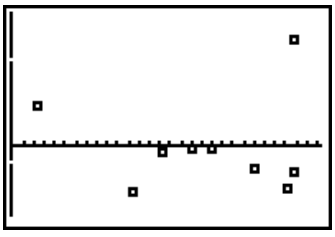


The next column contains the standard errors of each coefficient. These can be used to create confidence intervals for the true values using critical values for the t distribution for $n - p$ degrees of freedom. Finally we see the t statistics and p-values for testing H_0: $\beta_i = 0$ against H_A: $\beta_i \neq 0$. These suggest the coefficient of age is not significantly different from 0; in other words, mileage is a much more determining quantity for the price of used Corvettes which helps explain why its coefficient changed less than the coefficient of age from the single variable regressions.

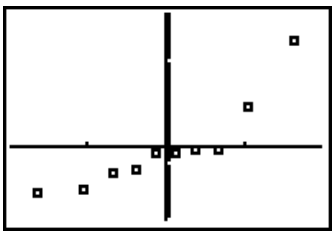
ASSESSING THE MODEL

Just as with simple (one-variable) linear regression, we will use residuals plots to assess the model. The program has stored fitted values (y-hats) in `L5` and residuals in `L6`.

Using the `STAT PLOT` menu, define a scatter plot of the residuals in `L6` (as Y) against the fitted values in `L5`. Pressing ZOOM 9 displays the plot. Just as with simple regression we are looking in this plot for indications of curves or thickening/narrowing which indicate problems with the model. With this small data set these are somewhat hard to see, but clearly the only positive residuals are for the largest and smallest fitted values, which could indicate a potential problem.



We define a normal plot of the residuals (as in Chapter 4). This normal plot is not a straight line, which indicates a violation of the assumptions. This multiple regression model is not appropriate for these data.



MULTIPLE REGRESSION CONFIDENCE INTERVALS

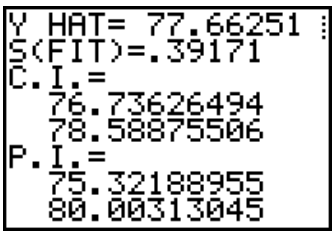
How well do the midterm grade and number of missed classes predict final grades? Data for a sample of students are below.

Final Grade, y	Midterm Exam, x_1	Classes Missed, x_2
81	74	1
90	80	0
86	91	2
76	80	3
51	62	6
75	90	4
48	60	7
81	82	2
94	88	0
93	96	1

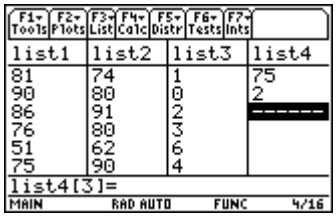
Performing the regression as above, we find the estimated regression equation $Final = 49.41 + 0.502 * Midterm - 4.71 * ClassesMissed$. All coefficients are significantly different from zero and a normal plot of the residuals is relatively straight. We'd like to use the model to create a prediction of the average grade for students with a midterm grade of 75 and 2 absences (a confidence interval); we also want to predict the grade for a particular student with a midterm grade of 75 and 2 absences (a prediction interval).

When executing the MULREG program, select `1:YES` in answer to the question `PREDICT Y`. You will be asked to input the independent variables' values. Enter them inside curly braces ([2nd][(] and [2nd][)]) separated by commas. You will then be asked for the desired confidence level.

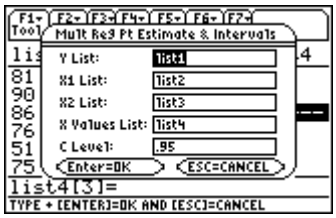
Pressing [ENTER] performs the calculations and displays the screen at right. We find the point estimate of the final grade for a student with a 75 midterm grade and two absences is 77.66, or 78 (practically speaking). The standard deviation of the fit at that point is 0.392. The confidence interval says we are 95% confident the average final grade for all students with a 75 midterm grade and two absences will be between 76.7 and 78.6. Further, we are 95% confident an individual student with a 75 midterm grade and two absences will earn a final grade between 75.3 and 80. Pressing [ENTER] at this point returns you to the Predict or Quit menu.



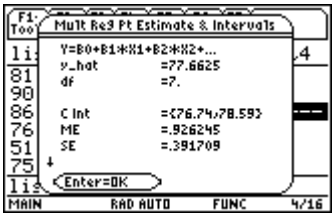
TI-89 Procedure



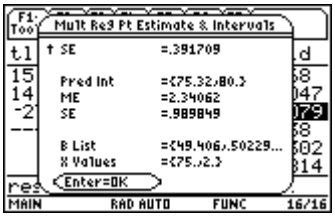
The data are entered with final grades in `list1`; midterm grades in `list2`, and absences in `list3`. Order of specification in the regression input screen matters. In `list4` we have entered the values for the predictions of interest. They must be entered in the order in which the xlists will be specified.



Press 2nd F2 (F7) and select option `8:MultRegInt`. We are first asked the number of independent variables. In our data we have 2. Use the right arrow to access the list of possible values, and select the appropriate one for your data set. Press ENTER to continue. Now specify the lists to be used in the regression and the list containing the values to be used in the intervals as at right. Specify the desired confidence level, here 95%, as a decimal.



Pressing ENTER performs the calculations and displays the screen at right. We find the point estimate of the final grade for a student with a 75 midterm grade and two absences is 77.66 or 78 (practically speaking). The confidence interval says we are 95% confident the average final grade for all students with a 75 midterm grade and two absences will be between 76.7 and 78.6.



Pressing the down arrow several times displays the remainder of the output. We are 95% confident an individual student with a 75 midterm grade and two absences will earn a final grade between 75.3 and 80. The first portion of the coefficients list and the X values used for the intervals are displayed as well.

WHAT CAN GO WRONG?

Not much that hasn't already been discussed. The most common errors are misspecification of lists and having more than one plot "turned on" at a time.

How do I get rid of those extra lists?

Press 2nd + (`MEM`). Select `2:Delete`, then select `4:List`. Arrow to the lists to be deleted and press ENTER.

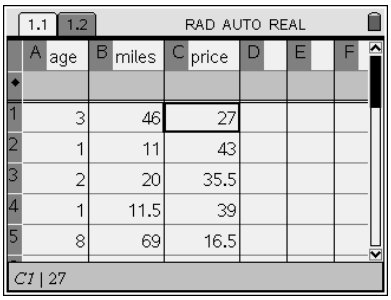
MULREG Program Listing. (The program can also be downloaded from the text's website.)

```
Disp "DATA IN COLS","OF [A]"
dim([A]) STO▶ θ1:Lθ1(1) STO▶ N:Lθ1(2) STO▶ L:L-1 STO▶ K
{N,1} STO▶ dim([B]):Fill(1,[B])
augment([B],[A]) STO▶ [B]
[B]^T[B] STO▶ [D]
seq([D](I,1),I,2,L+1)/N STO▶ Lθ1
{L,L} STO▶ dim([C]):Ans STO▶ dim([E])
{N,1} STO▶ dim([C])
Input "RESPONSE COL=",R
For(I,1,N):[A](I,R) STO▶ [C](I,1):End
If R≤K:Then
For(J,R+1,L):For(I,1,N)
[B](I,J+1) STO▶ [B](I,J)
End:End
[B]^T[B] STO▶ [D]
End
{L,L} STO▶ dim([D]):[D]x⁻¹ STO▶ [D]
{N,L} STO▶ dim([B])
[B]^T [C]:[D]Ans STO▶ [E]
Matr▶list([E],Lθ1)
Disp "COEF Lθ1=":Pause Lθ1
[B][E] STO▶ [B]
Matr▶list([C],LY)
Matr▶list([B],LYP)
mean(LY) STO▶ Y
sum((LY-Y)^2) STO▶ T
sum((LY-LYP)^2 STO▶ E
LYP STO▶ L5
LY-LYP STO▶ L6
DelVar LY:DelVar LYP
N-L STO▶ M:T-E STO▶ R:R/K STO▶ Q:E/M STO▶ D:√(D) STO▶ S:Q/D STO▶ F
S√(seq([D](I,I),I,1,L)) STO▶ Lθ2
Disp "STDEV Lθ2=":Pause Lθ2
Lθ1/Lθ2 STO▶ Lθ3
Disp "T-RATIO Lθ3=":Pause Lθ3
Disp "COEF P Lθ4="
1-2seq(tcdf(0,abs(Lθ3(I)),M),I,1,L) STO▶ Lθ4
Pause Lθ4
1-(N-1)*D/T STO▶ A
ClrHome
Output(1,1,"S="
Output(1,9,S
Output(2,1,"R^2="
Output(2,9,R/T
Output(3,1,"R^2ADJ="
Output(3,9,A
Output(4,1,"REG DF="
Output(4,9,K
Output(5,1,"ERR DF="
Output(5,9,M
Output(6,1,"TOT DF="
Output(6,9,N-1
```

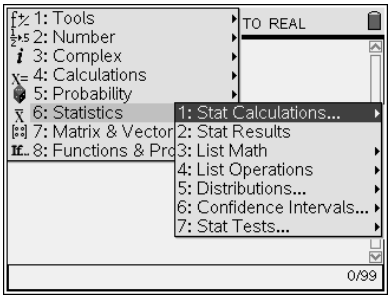
```
Pause
ClrHome
Output(1,1,"SS REG="
Output(1,9,R
Output(2,1,"SS ERR="
Output(2,9,E
Output(3,1,"SS TOT="
Output(3,9,T
Output(4,1,"MS REG="
Output(4,9,Q
Output(5,1,"MS ERR="
Output(5,9,D
Output(6,1,"F=    "
Output(6,9,F
Output(7,1,"P-VAL="
1-Fcdf(0,F,K,M) STO▶P
round(P,5) STO▶P
Output(7,9,P
Pause
ClrHome
Lbl C
Menu("PREDICT Y?","YES",A,"QUIT",B)
Lbl A
Input "X{}=",Lθ5
Input "CONF. LEVEL=",C
{L,1}STO▶dim([C])
1STO▶ [C](1,1)
For(I,1,K):Lθ5(I) STO▶ [C](I+1,1):End
[E]^T[C]
Ans(1,1) STO▶Y
ClrHome
round(Y,5) STO▶Y
Output(1,1,"Y HAT="
Output(1,8,Y
[C]^T[D][C]
Ans(1,1) STO▶V
round(S√(V),5) STO▶T
Output(2,1,"S(FIT)="
Output(2,8,T
TInterval 0,√(M+1),M+1,C
upperSTO▶T
Output(3,1,"C.I.="
Y+TS√(V) STO▶D
Y-TS√(V) STO▶E
Output(4,3,E
Output(5,3,D
Output(6,1,"P.I.="
Y-TS√(1+V) STO▶E
Y+TS√(1+V) STO▶D
Output(7,3,E
Output(8,3,D
Pause
Goto C
Lbl B
DelVar [C]:DelVar [D]:DelVar [E]:DelVar Lθ5
ClrHome
```

Commands for the TI-Nspire™ Handheld Calculator

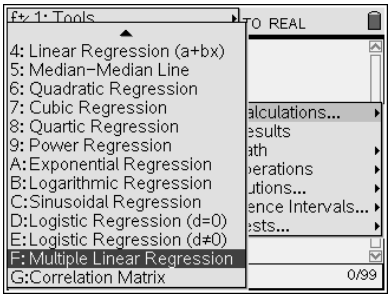
To conduct multiple regression, first place the data into lists. Values for the automobiles example are shown.



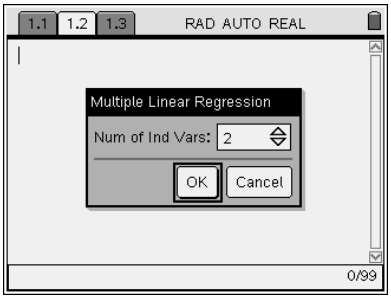
To compute the regression equation, start on a Calculator page. Press (menu), select Statistics and Stat Calculations



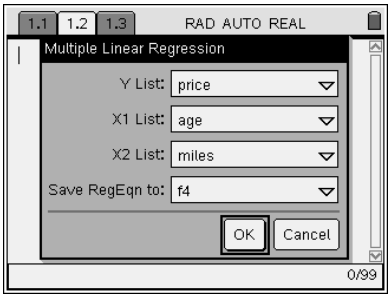
Select Multiple Linear Regression.



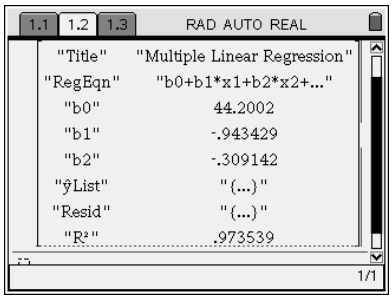
Select the number of independent variables; two for this example.



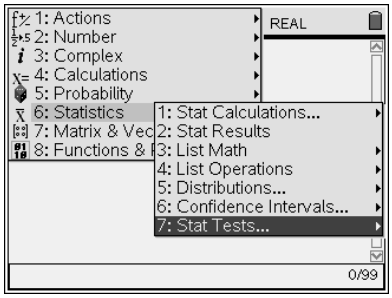
Select the list names. Be sure to select the *y* variable for the first input.



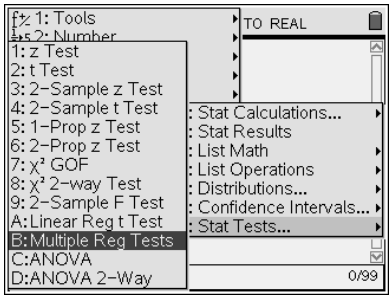
The negative coefficients predict that price will decrease as age increases and mileage increases.

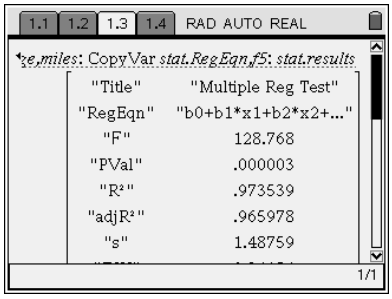


To conduct a hypothesis test, open a Calculator page. Press (menu) , select Statistics, and then Stat Tests.

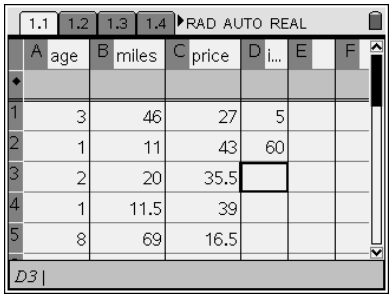


Select Multiple Reg Tests. Select the number of independent variables as shown above. Then select the list names, also as shown above. The *F* Statistic, *p*-value, R^2 and other information is displayed.

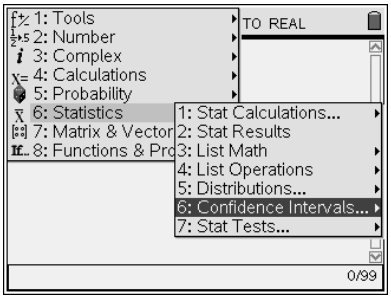




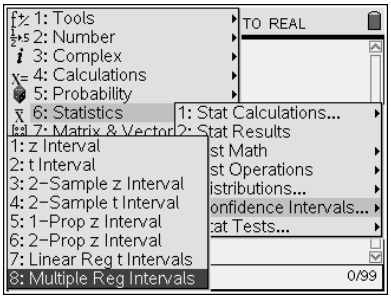
To conduct a confidence interval, we need one additional list that shows the values for the independent variables for which we are predicting. In this example we need an age and the number of miles to predict the price. For example, the list *ind* shows *age* = 5 and *miles* = 60.



To actually construct the interval, open a Calculator page. Press (menu) , select Statistics, and then Confidence Intervals.



Select Multiple Reg Intervals. Select the number of independent variables as shown above.



Select the list names, including the fourth list for the independent variable values and the confidence level.