EXPLORATIONS

Chapter 4

Exploring Bivariate Data

Topic 10 covers scatterplots, Topic 11 discusses correlation and least-squares regression, and Topic 12 covers transformation to achieve linearity.

Topic 10—Scatterplots

Example: Use the list of building heights from Topic 1 (list **phily**). Use the list of building completion dates from Topic 2 (list **yrphil)**.

Scatterplots

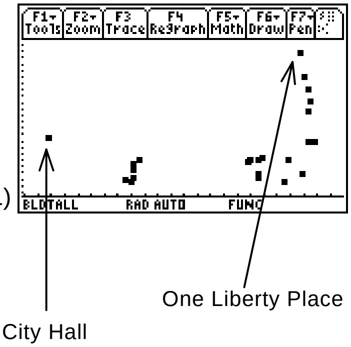
In the Stats/List Editor, delete list **seattle** and replace it with the list name **yrphil** from Chapter 1, Topic 2. Delete the last value in list **phily**, **417** feet, by highlighting it and pressing [◆] [DEL], as the year it was completed is unknown. List **yrphil** is in order with the last, 23rd value, being 1992. Delete the 24th value if there is one in your list.

1. From the Stats/List Editor, turn off all plots with [F2] **Plots**, **3:PlotsOff**.
2. Press [F2] **Plots**, **1:Plot Setup** and define **Plot 1** as Plot Type: **Scatter**, Mark: **Square**, X List: **yrphil**, and Y List: **phily**.
3. From the Plot Setup screen, press [F5] **ZoomData** (screen 1).
4. Press [F3] **Trace** and the arrow keys [▶] and [◀] to help identify some of the clusters.

The first building on the left is City Hall Tower, completed in 1901 and measuring 585 feet to the top of the hat on the statue of William Penn.

The next cluster of six buildings was built in the late 1920s and early 1930s. These are followed by another cluster of six in the late 1960s (1969) and early 1970s. There are two in the early 1980s (1982 and 1983).

The tallest building (945 feet) was built in 1987 (One Liberty Place) and until that time no building was taller than City Hall. Since then four other buildings have been built that are taller than City Hall.

(1)


Note: *If you get a dimension mismatch error, make sure there are 23 values in each list.*

Using Multiple Symbols to Identify Parts of List

To highlight these few buildings:

1. In the Stats/List Editor, clear **list1** and **t1** and enter these new values:

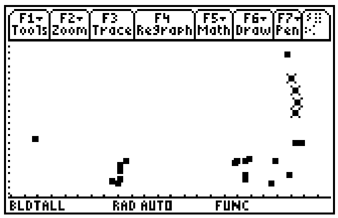
   list1: 1989, **1990**, **1990**, **1991**

   t1: 848, **700**, **792**, **739**

2. Set up and define **Plot 2** as Plot Type: **Scatter**, Mark: **Cross**, X List: **list1**, and Y List: **t1**.

3. Press [ENTER] twice and then press [◆] [GRAPH] (screen 2).

The four tallest buildings (since City Hall) completed after 1987 are now designated with × at the upper right of screen 2.

(2)


Note: *In Topic 11, you will use a category list to give different parts of a scatterplot different symbols.*

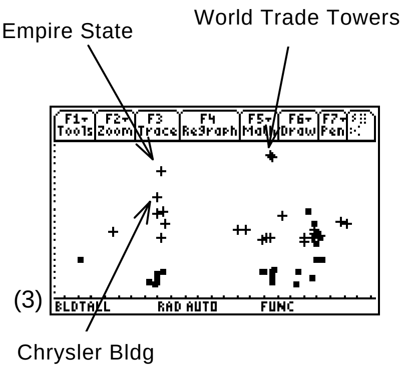
Comparing Two Scatterplots

Example: Tall buildings in Philadelphia and New York City (saved in Topic 9).

Store (in the order given) the completion dates for the 24 tallest buildings in New York City (provided below) in list **yrnyc**.

1913	**1930**	**1930**	**1931**	**1931**	**1932**	**1933**	**1960**	**1963**	**1969**	**1971**	**1972**
1972	**1973**	**1977**	**1985**	**1985**	**1988**	**1989**	**1989**	**1989**	**1991**	**1999**	**2001**

1. With the **phily** data set in **Plot 1** above, define **Plot 2** as Plot Type: **Scatter**, Mark: **Plus**, X List: **yrnyc**, and Y List: **nyc**.
2. With both **Plot 1** and **Plot 2** selected from the Plot Setup screen, press F5 **ZoomData** (screen 3).



In addition to many buildings taller (higher on the screen) than those in Philadelphia, you also see a cluster in the early 1930s including the Chrysler and Empire State Buildings. There is a big gap after the Depression, then two built in the early 1960s, then another cluster, like Philadelphia, in the late 1960s and early 1970s with the two tallest buildings completed in 1972 and 1973. Were there tax incentives for building office space during those years? How did new technology play a part in the construction of buildings? There is a pattern and a story to what might seem like random points on a screen.

The following are the years that the first 17 tallest buildings in list **seattle** (Topic 9) were completed, in case you wish to investigate that data in the manner as described above.

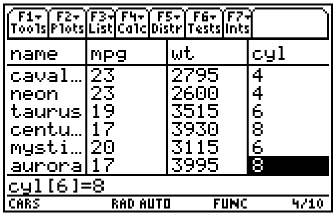
1914	**1962**	**1969**	**1973**	**1976**	**1977**	**1980**	**1981**	**1981**
1981	**1983**	**1985**	**1986**	**1988**	**1989**	**1989**	**1990**	

Topic 11 will also explore patterns in a scatterplot with another data set.

Topic 11—Correlation and Least-Squares Regression Line

Example: City gas mileage, in miles per gallon (mpg), for some automatic transmission cars of known weight (wt) in pounds with a known number of cylinders (cyl). The data is given in the following table.

Create a folder called **CARS** (as shown in Topic 1, *Creating a New Folder* section). Change to folder **CARS** (as shown in Topic 1, *Changing Folders While in the Stats/List Editor* section) and store the data given below in lists: **name**, **mpg**, **wt**, and **cyl** as partially shown in screen 4.

(4)


Note: *Variable names are limited to eight characters, so you might choose to enter either the make or the name of the car.*

Name	MPG	WT	CYL
Chevrolet (Cavalier)	23	2795	4
Dodge (Neon)	23	2600	4
Ford (Taurus)	19	3515	6
Lincoln (Centurion)	17	3930	8
Mercury (Mystique)	20	3115	6
Olds (Aurora)	17	3995	8
Pontiac (Grand Am)	22	3115	4
Cadillac (Deville)	17	4020	8
Chrysler (Sebring)	19	3175	6
BMW 3-Series (BMW3S)	19	3225	6
Ford (Crown Victoria)	17	3985	8
Mazda (Protégé)	29	2500	4
Hyundai (Accent)	28	2290	4

(Source: M. Triola, *Elementary Statistics,* 8th edition (page 803), © 2001 Addison Wesley Longman Inc. Reprinted by permission of Addison Wesley Longman.)

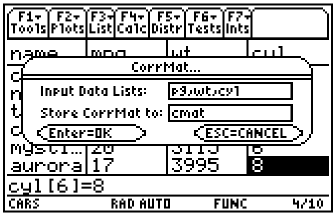
Correlation Matrix

A correlation matrix is an array of row-column entries, each of which represents a coefficient of correlation, *r*. The $r_{x,y}$ entry is the coefficient of correlation between the variable in the *x* row and the *y* column. For example, if **mpg** is row 2 and **wt** is column 1, then the coefficient of correlation between **mpg** and **wt** is found by reading the entry in the second row, first column of the matrix.

Which variable is most highly correlated with mpg?

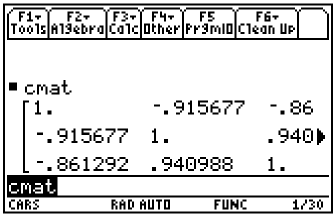
1. In the Stats/List Editor, press [F4] **Calc**, **5:CorrMat**, with Input Data Lists: **mpg**, **wt**, and **cyl**, and the results to be stored in folder **CARS** as **cmat**, and then press [ENTER] (screen 5).
2. Press [ENTER] to display **Done**.

(5)



3. On the Home screen, type **cmat** and then press [ENTER] (screen 6).

(6)



The simple linear correlation coefficient between **mpg** and **wt** is *r* = -0.915677, and between **mpg** and **cyl** is *r* = -.861292 (going down the first column or across the first row of the correlation matrix in screen 6).

Also notice that the correlation coefficient between **cyl** and **wt** is *r* ≈ +0.94. The greater number of cylinders tends to go with the heavier cars (positive correlation). The heavier the car (or number of cylinders), the smaller the number of miles per gallon (negative correlation of -.92 (or -0.86)).

Note: *For larger matrices, see Topic 9, screen 7 on how to scroll up or down, left or right on the Home screen.*

Scatterplots and Linear Correlation Coefficients

The previous correlation coefficients indicate that the plot of the two variables should be fairly linear, |*r*| is close to 1. If you were to compute the correlation coefficient between New York City building heights and the year they were completed (pluses in Topic 10, screen 3), then *r* = -0.23. The correlation coefficient between Philadelphia building heights and the year they were completed would be computed as *r* = 0.45 (Topic 10, screen 1).

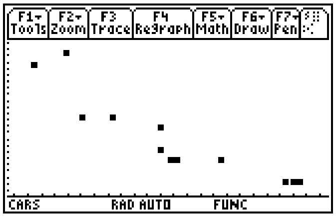
To show the relationship between weight and miles per gallon:

1. From the Plot Setup screen:

   a. Deselect **Plot 2** from the previous exercise.

   b. Define **Plot 1** with Plot Type: **Scatter**, Mark: **4:Square**, X List: **wt**, Y List: **mpg**, and Use Freq and Categories?: **NO**.

   c. From the Plot Setup screen, press F5 **ZoomData** (screen 7). Screen 7 confirms the inverse relationship indicated by the correlation coefficient for *r* = -0.92 (screen 3).

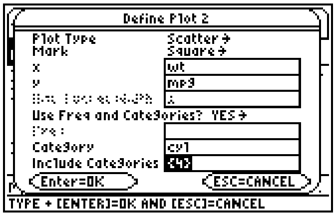
(7)


2. To show the effect of the number of cylinders:

   a. Deselect **Plot 1**.

   b. Define **Plot 2** with Plot Type: **Scatter**, Mark: **4:Square**, X List: **wt**, Y List: **mpg**, Use Freq and Categories?: **YES**, Category: **cyl**, and Include Categories: **{4}** (screen 8). (**4** represents 4-cylinder cars.)

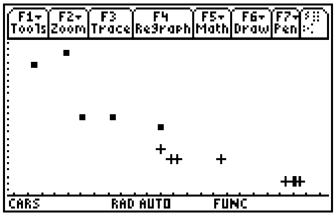
   c. Press ENTER.

(8)


   d. Define **Plot 3** like **Plot 2**, but with two differences: Mark: **Plus** and Include Categories: **{6,8}**. (**6,8** represents 6- and 8-cylinder cars.)

   e. Press ◆ [GRAPH] (screen 9).

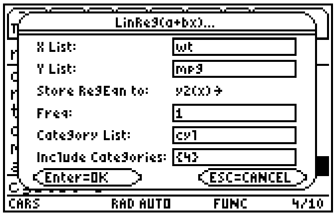
You notice the 4-cylinder cars tend to be lightest with better (greater) gas mileage, while the 6- and 8-cylinder cars are heavier and get poorer gas mileage.

(9)


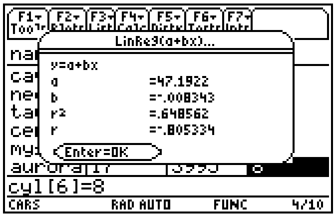
3. To compare the correlation coefficient for the two categories, use groupings of 4, 6, and 8 cylinders:

   a. In the Stats/List Editor, press F4 **Calc**, **3:Regressions**, **1:LinReg(a+bx)**.

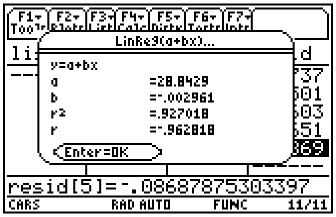
b. Define X List: **wt**, Y List: **mpg**, Store RegEqn to: **y2(x)**, Freq: **1**, Category List: **cyl**, and Include Categories: **{4}** (screen 10).

(10) 

c. Press [ENTER] [ENTER]. Observe that the coefficient of correlation is r = -0.81 (screen 11).

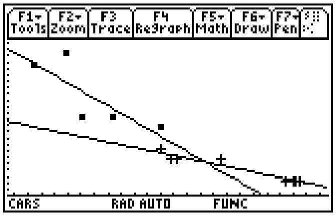
(11) 

d. Repeat steps **3a** – **c** above, but with two changes: Store RegEqn to: **y3(x)** and Include Categories: **{6,8}**. Observe that the coefficient of correlation this time is r = -0.96 (screen 12).

(12) 

Least-Squares Regression Lines

To show the least-squares regression lines for the above two categories, after steps 2 and 3 in the previous section, press [◆] [GRAPH] (screen 13).

(13) 

Screen 13 graphically confirms the results of screens 11 and 12. The pluses (6 and 8 cylinders) are much closer to their least-square regression line (with r = -0.96) with a smaller slope (b = -.003), than the squares (4 cylinders) are to their regression line (with r = -.81) with a steeper slope (b = -.008).

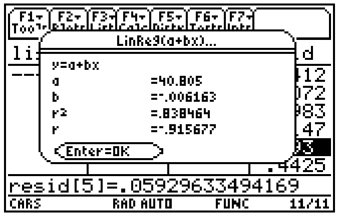
2. To plot the least-squares regression line through all the data, you must first turn off the other plots and regression lines.

   a. From the Stats/List Editor, turn off all functions and plots with [F2] **Plots**, **4:FnOff** and [F2] **Plots**, **3:Plots Off**.

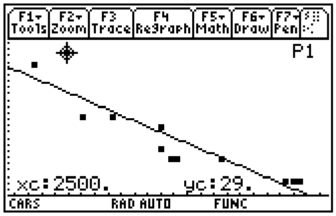
   b. From the Plot Setup screen, select **Plot1** by pressing [F4] (√), and then press [F5] **ZoomData**.

c. In the Stats/List Editor, press [F4] **Calc**, **3:Regressions**, **1:LinReg (a + bx)** with X List: **wt**, Y List: **mpg**, Store RegEqn to: **y1(x)**, Freq: **1**, and both Category List: and Include Categories: cleared.

d. Press [ENTER] and observe that r = -0.915677 (as in screen 6) and the regression line y = 40.805 -0.006163x (in the form $y = a + bx$) (screen 14).

(14) 

e. Press [◆] [GRAPH], [F3] **Trace**, and the ▶ key eleven times (screen 15).

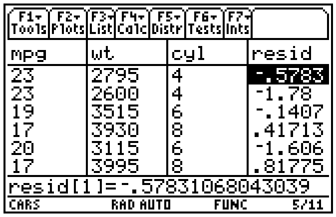
(15) 

The highlighted point is far from the regression line. This is the 2500-pound, 4-cylinder Mazda Protégé that gets 29 mpg in the city.

It also appears that an upward concave curve might better fit the data than the linear regression line.

Residual Plots and Outliers

When you return to the Stats/List Editor, note the new list, **resid**, (pasted at the end of the editor), with the first value -0.5783 ≈ -.58 (screen 16).

(16) 

Using the regression line with **x = 2795 lbs** for **wt** and **y = 23 mpg** (screen 16), calculate the predicted value, y.

y = 40.805 – 0.006163 * 2795 ≈ 23.58 mpg.

Now calculate the difference in the observed and the predicted value, y.

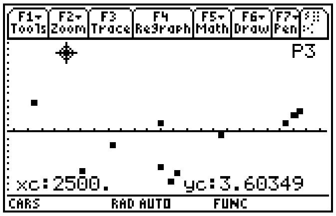
y-y = 23 - 23.58 = -0.58 to get the residual in the first row.

1. From the Stats/List Editor, turn off all functions and plots with [F2] **Plots**, **4:FnOff** and [F2] **Plots**, **3:Plots Off**.

2. Set up and define **Plot 3** as Plot Type: **Scatter**, Mark: **4:Square**, X List: **wt**, Y List: **resid** (must be pasted from folder **STATVARS** to carry along the path designation), and Use Freq and Categories?: **NO**.

3. From the Plot Setup screen, press F5 **ZoomData** (screen 17).

4. Press F3 **Trace** and the arrow keys to reveal the largest positive residual of 3.6 mpg for the Mazda Protégé in the upper left corner, and the largest negative residual of -2.2 mpg for the 3,175 lb., 6-cylinder Chrysler Sebring.

(17)


Consider these both possible *outliers*, since they are far from the regression line. In Topic 42, you will discuss how far is far enough to be considered an outlier.

Influential Points

Since the Mazda Protégé is far from the regression line and far from the center of the data indicated by point $(\bar{x}, \bar{y})$, it could have an influence on the regression line. Check this by plotting the data without this point, but first repeat screen 15 for all the points.

1. To repeat step 2 (screens 14 and 15):

   a. From the Stats/List Editor, turn off all functions with F2 **Plots**, **4:FnOff**.

   b. Press F4 **Calc**, **3:Regression**, **1:LinReg(a+bx)**, with X List: **wt**, Y List: **mpg**, Store RegEqn to: **y1(x)**, Freq: **1**, and both Category List: and Include Categories: cleared.

   c. Press ENTER ENTER.

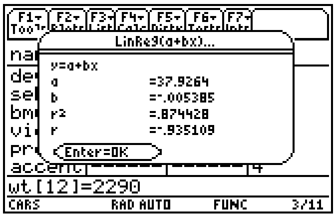
   d. Press F2 **Plots**, **1:Plot Setup** and deselect all but **Plot 1**.

   e. From the Plot Setup screen, press F5 **ZoomData**, but do not trace (screen 15).

2. In the Stats/List Editor, delete the 12th value in both lists **mpg** (**29**) and **wt** (**2500**).

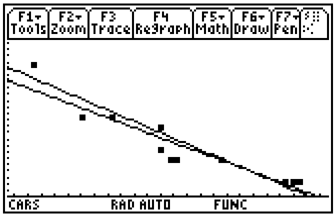
3. Press F4 **Calc**, **3:Regression**, **1:LinReg(a+bx)**, with X List: **wt**, Y List: **mpg**, Store RegEqn to: **y2(x)**, Freq: **1**, and Category List: cleared.

4. Press ENTER (screen 18).

(18)


Comparing screen 18 and screen 14 shows *r* has increased in magnitude, indicating the points will be closer to the regression line. The slope changes from $b = -0.006163$ to $b = -0.005385$ so the slope is less steep.

5. Press [◆] [GRAPH] (screen 19). The lower line (without the Mazda Protégé) was influenced by the exclusion of the Mazda Protégé.

(19) 

With all values but the Chrysler Sebring, you would obtain $y = 41.1 - 0.0062x$, $r^2 = 0.86$, $r = -0.93$, compared to $y = 40.8 - 0.0062x$, $r^2 = 0.84$, $r = -0.92$ in screen 14.

This shows some improvement in the points being close to the regression line, but with no noticeable change in the regression line, so the Chrysler Sebring is not an influential point.

In Topic 12, you will use the same data to predict the gas mileage in mpg given a car's weight in pounds. You will also discuss how to transform the data to make better predictions.

Topic 12—Transformation to Achieve Linearity

Example: Weights and gasoline mileage data for cars as in Topic 11. It is assumed you are familiar with the procedure and notation of Topic 11.

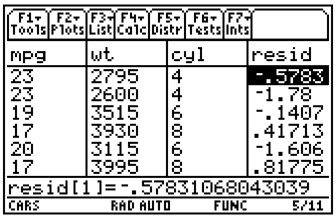
Linear Least-Square Fit and the Coefficient of Determination r^2

From Topic 11, screen 14, $r^2 = 83.8\%$,
mpg = 40.805 - .006163 **wt**. This includes the Mazda Protégé data, so make sure that this data point is in the lists **mpg**, **wt**, and **cyl** by re-entering the 12th value from the table. See Topic 1, Steps 3 through 5 in the *Editing a List in the Stats/List Editor* section.

Repeat these results from the Stats/List Editor:

1. Press [F4] **Calc**, **3:Regressions**, **1:LinReg (a + bx)**, with X List: **wt**, Y List: **mpg**, Store RegEqn to: **y1(x)**, and Freq: **1**.

After returning to the Stats/List Editor, notice the first two values in list **resid** are now **res[1] = -.5783** and **resid [2] = -1.78** as in screen 20 and in Topic 11, screen 16.

(20) 

Note: *You need to make sure the* **resid** *list has the correct values before the following calculations can be done properly.*

For the next few topics, change the MODE to **Approximate**.

1. Press [MODE] and then press [F2] **Page2**.

2. Press [⊙] to select Exact/Approx and select **3:APPROXIMATE**.

3. Press [ENTER] [ENTER].

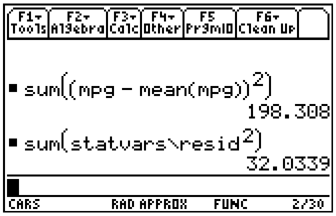
4. From the Home screen, calculate the sum of squares about the mean for y, or the sum of squares total: SST = $\Sigma(y - \bar{y})^2$ = 198.308.

   Type: **sum((mpg - mean(mpg))^2)** (first line of screen 21).

5. Calculate the sum of square error:

   $$\text{SSE} = \Sigma(y - \hat{y})^2 = 32.0339.$$

   Type: **sum(statvars\resid^2)** (second line of screen 21).



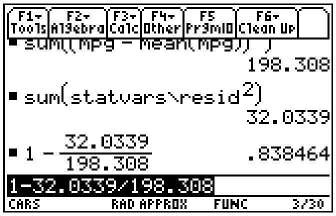
(21)

Note: *On all calculations in steps 4 through 7, you can use the* CATALOG *key to paste* **sum** *and* **mean** *if you prefer not to type. You can paste* **mpg**, **wt**, *and* **statvars\resid** *from* [VAR-LINK].

6. Calculate r^2:

   r^2 = 1 – SSE/SST =0.838464 = 83%.

   Type: **1 – 32.0339/198.308** (third line of screen 22).

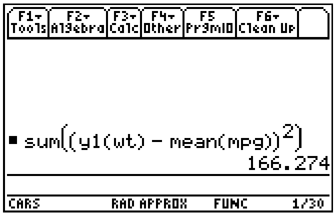


(22)

7. Calculate the sum of squares regression:

   $$\text{SSR} = \Sigma(\hat{y} - \bar{y})^2 = 166.253.$$

   Type: **sum((y1(wt) – mean(mpg))2)** (screen 23).



(23)

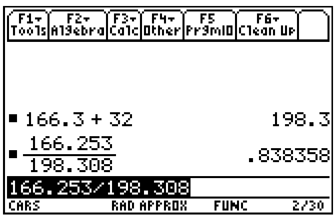
You will use r^2 as a measure of how well you have achieved linearity.

Since SST = SSR + SSE, then SSR = SST – SSE. You can also find r^2 from SSR and SST:

$$r^2 = 1 - \frac{\text{SSE}}{\text{SST}} \text{ (from step 6).}$$

Getting a common denominator, you have

$$\frac{\text{SST} - \text{SSE}}{\text{SST}} = \frac{\text{SSR}}{\text{SST}} = \frac{166.253}{198.308} = .838 = 83.8\% \text{ (screen 24).}$$



(24)

This is always true for least-squares linear regression. 83.8% of the deviations about the mean are explained by the regression line.

If SSE = 0, all points are on the line and r^2 = 100%. If $\hat{y} = \bar{y}$ (a constant of zero slope) then SSE = SST and r^2 = 0.

The equation **mpg** = 40.805 - .006163 (**wt**) from Topic 11, screen 14 is a least-square line fit in that no other straight line through these data points will give an SSE less than 32.03.

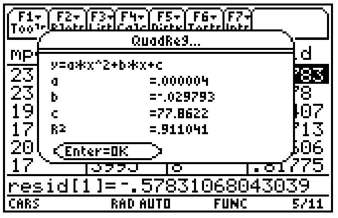
Quadratic Regression

Although you will not transform the data to do this regression, it is helpful for what will follow. The motivation for this regression was the parabolic shape of the data points and of the residuals in Topic 11, screens 17 and 19.

From the Stats/List Editor:

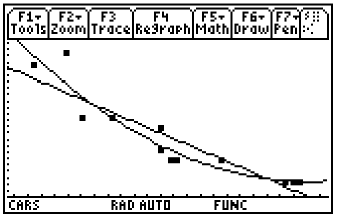
1. Press [F4] **Calc**, **3:Regressions**, **4:QuadReg**, with X List: **wt**, Y List: **mpg**, Store RegEqn to: **y4(x)**, and then press [ENTER] (screen 25).
2. From screen 25, r^2 **= 91.1%** and **mpg = .000004**x^2**- 0.029793**x **+.77.8622**.
3. Press [ENTER] to return to the Stats/List Editor screen with list **resid**.

   resid[1] = -.0736 and **resid[2] = -2.047**.

(25) 

Note: *r^2 has increased from* **83.8%** *to* **91.1%***; the first residual decreased in magnitude but the second increased in magnitude.*

4. From the Plot Setup screen, define **Plot 1** with Plot Type: **Scatter**, Mark: **4:Square**, X List: **wt**, and Y List: **mpg**.
5. Deselect all functions except **y1** and **y4** by pressing ◆ [Y=].
6. From the Plot Setup screen, press [F5] **ZoomData** (screen 26).

(26) 

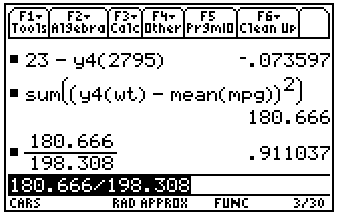
Note*:* **y1(x) = 40.805 - 0.006163x** *from Linear Regression in Topic 11, screen 14, while* **y4(x) = 77.8622 - 0.029793**x **+ .000004**x^2 *from Quadratic Regression (screen 25).*

7. Pressing [◆] [Y=] reveals **y4(x) = 3.646143 E -6·x^2 - 0.029793x + 77.862163** with more significant figures than screen 25.

8. On the Home screen, type **23 - y4(2795)** and press [ENTER] to display **-.073597**, the first residual (first line of screen 27).

   Type: **sum((y4(wt) - mean(mpg))2)** as in the second line of screen 27.

   Observe that the SSR value is 180.666. SST = 198.308 from before, so $r^2 = \frac{\text{SSR}}{\text{SST}} = \frac{180.666}{198.308}$ = 91.1% (screen 27).

(27) 

Note: **SSE = SST – SSR = 17.6 = sum (statvars\resid²).**
$y = c + bx + ax^2 = c + bx_1 + ax_2$ is linear in the variables a, b, and c with $x_2 = x_1^2$.

Quadratic regression is a least-squares fit in that the SSE is the smallest possible when fitting a parabola to the data.

Note: *CubicReg and QuartReg are other polynomial regressions built into the TI-89. CubicReg does not give a noticeably better fit for the above data.*

Transformations that Linearize

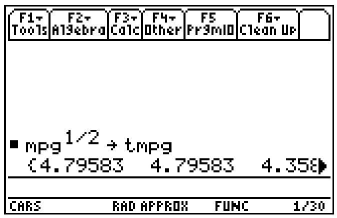
Taking a power of ***y*** or ***x*** (or both) can change the shape of the data. For example, $100^{1/2} = \sqrt{100} = 10$ while $\sqrt{1} = 1$. Since your data curves up, you can bring larger ***y*** values down further than smaller values by taking their square root or other decreasing powers.

Note: *If the shape of the data were curved down, you might want to raise some values quickly by raising values by a positive power, for example,* $10^2 = 100$, $1^2 = 1$.

Example: $\sqrt{y} = y^{1/2}$, Square Root Transformation.

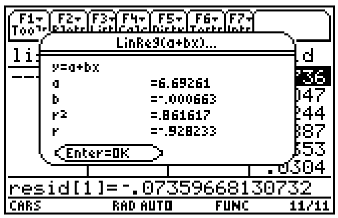
1. From the Home screen, type **mpg^(1/2)→tmpg** and press [ENTER] to display **{4.79583…, }** (screen 28).

   This is a list of square roots of **mpg** values. For example, the first value in list **mpg** is 23. In screen 28, the first value displayed is 4.79583, which is $\sqrt{23}$.

(28) 

Note: **tmpg** *for transformed* **mpg**.

2. From the Stats/List Editor, press [F4] **Calc**, **3:Regression**, **1:LinReg (a+bx)**, with X List: **wt**, Y List: **tmpg** (in folder **CARS**), and Store RegEq to: **y2(x)** (screen 29).

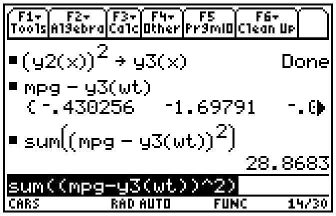
(29) 

r^2 **= 86.2%** and $\sqrt{\text{mpg}} = 6.69261 - 0.000663\ (\text{wt})$ or $\text{mpg} = (6.69261 - 0.000663\ \text{wt})^2$. To check the first residual to see how close the first transformed data point is to the fitted straight line, calculate:

$\sqrt{23} - (6.69261 - 0.000663 * 2795) = -.04$. This agrees with the first value in the transformed **resid** list.

3. From the Home screen, type **(y2(x))^2→y3(x)**, and press ENTER to display the top line of screen 30.

4. Enter **mpg - y3(wt)** and press ENTER to display **{-.430256 -1.69791 …}**, the second calculation in screen 30. This is how close the fitted curve is to the original data, the residuals associated with this quadratic model. The first and second residuals ≈ **-.43** and **-1.70**.

(30) 

Note: **y3(x) = (6.69261 - 0.000663x)^2**

5. Type **sum((mpg-y3(wt))^2)** and press ENTER to display **28.8683,** the third calculation (screen 30).

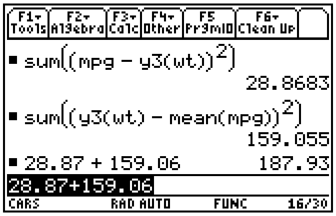
This is the sum of the squares error, SSE.

Note that $1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{28.8683}{198.303} = .854$, which is different from the r^2 value of .862 that you obtained in step 2.

r^2 measures only how well the straight line fits the transformed data.

6. Type: **sum((y3(wt) - mean(mpg))^2)** and press ENTER to display 159.055, the second calculation in screen 31.

This represents the sum of squares residuals, SSR. Adding SSE + SSR, you obtain 28.87 + 159.06 = 187.93 (third calculation in screen 31). Note that 187.93 ≠ 198.308 which was the SST from screen 21, so it makes no sense to talk about the percent explained by the regression curve.

(31) 

To summarize and to extend to other transformations:

$y = a + bx$ Linear Least-Squares Fit transformation (screens 20 and 21).

r^2 = 83.8%, SSE = 32.0, first and second residuals = -0.5783, -1.78, mpg = 40.805 - 0.006163 (wt).

$\sqrt{y} = y^{1/2}$ Square Root transformation (screens 28 to 31).

r^2 = 86.2%, SSE = 28.9, first and second residuals = -0.430, -1.70

$\sqrt{\text{mpg}} = 6.69261 - 0.000663\ (\text{wt})$ or $\text{mpg} = (6.69261 - .000663\text{wt})^2$

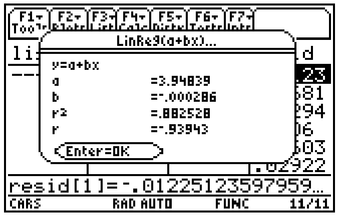
Other transformations:

$y^{1/3}$, cube root linear transformation (decreasing the exponent of y) gives r^2 = 86.9%, SSE = 28.0, first and second residuals = -0.38, -1.67.

ln(y) linear transformation:

Repeat steps 1 and 2 corresponding with screens 28 and 29, except use **ln(mpg) → tmpg** as the transformation statement.

As you see in screen 32, r^2 = 88.2528%, so the ln linear transformation equation is: ln(mpg) = 3.94839 - .000286 (wt).

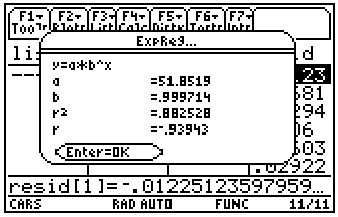


(32)

ExpReg transformation:

Press [F4] **Calc**, **3:Regressions**, **8:ExpReg**, with X List: **wt**, Y List: **mpg** (screen 33), with **y = 51.8519 * 0.999714^x**, and r^2 **= 88.2528%**.

Therefore, mpg = 51.852 $(0.999714)^{wt}$ is the exponential transformation equation.



(33)

The Stats/List Editor now has two lists pasted at the end: **resid** and **residt**.

The first two values of **resid** are **-.28** and **-1.62**, which are the residuals calculated from the final model as above.

The first two values of **residt** are **-0.0123** and **-.0681**, which are the residuals of the transformed straight line **ln(mpg) = tmpg = 3.94839 - .000286 (wt)** for which the r^2 **= .882528** was calculated.

y^{-2} linear transformation (decreasing the exponent of y so it is now negative).

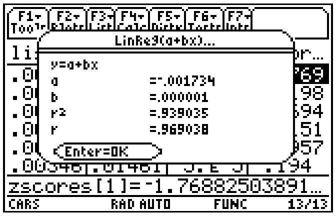
Note: $\mathbf{y^{-1}}$ *gave* r^2 **= 91.6, SSE = 22.54**. $\mathbf{y^{-3}}$ *gives very small numbers, for example* $\mathbf{23^{-3}}$ **= .00008** *and causes numerical problems, so stop at* y^{-2} *in this progression.*

From the Home screen, type: **mpg^-2 → tmpg**.

From the Stats/List Editor, press [F4] **Calc**, **3:Regressions**, **1:LinReg (a+bx)**, with X List: **wt**, Y List: **tmpg**, and Store RegEq to: **y2(x)**.

Observe the results in screen 34, with:

r^2 = 93.9%, so $(\text{mpg})^{-2} = -0.001734 + 0.0000013\ (\text{wt})$, or solving for mpg: $\text{mpg} = (-0.001734 + 0.0000013\ (\text{wt}))^{-1/2}$ is the y^{-2} linear transformation equation.



(34)

From the Home screen, type: **(y2(x)^(-1/2))→y3(x)** and then press [ENTER]. Now type: **mpg – y3(wt)→residt** and press [ENTER].

1. Press [◆] [Y=] (screen 35).

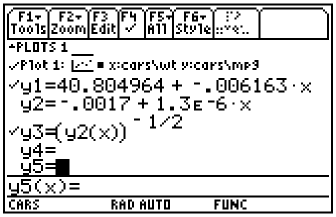
   y1 is the original linear fit, **y2** is this transformation linearization, while **y3** is the resulting fit.

2. Use [F4] (√) to be sure only **y1** and **y3** are selected (screen 35).

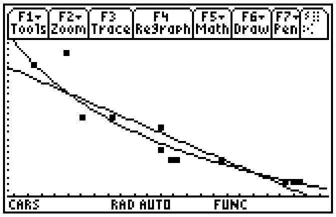
3. From the Stats/List Editor, press [F2] **Plots**, **1:Plot Setup**, and highlight **Plot 1**.

4. Press [F1] **Define**, with Plot Type: **Scatter**, Mark: **Square**, X List: **wt**, Y List: **mpg**, and Use Freq and Categories? **NO**.

5. On the return to the Plot Setup screen, press [F5] **ZoomData** (screen 36), which is similar to the quadratic fit in screen 26.



(35)



(36)

To look at a residual plot:

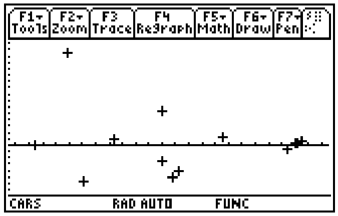
1. From the Stats/List Editor, turn off all functions and plots with [F2] **Plots**, **4:FnOff** and [F2] **Plots**, **3:PlotsOff**.

2. Press [F2] **Plots**, **1:Plot Setup**, and highlight **Plot 2**.

3. Press [F1] **Define**, with Plot Type: **Scatter**, Mark: **Plus**, X List: **wt**, and Y List: **residt**, and Use Freq and Categories?: **NO**.

4. From the Plot Setup screen, press [F5] **ZoomData** (screen 37).

   Other than a residual of 3.59 in the upper left corner (the Mazda Protégé) this plot looks fairly random.

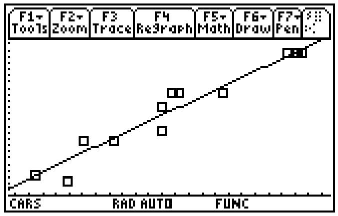
(37)


Note: *Randomness of residuals is desired to make confidence and prediction intervals in Topics 33 and 42.*

You can plot the transformed data to see how well it fits a straight line:

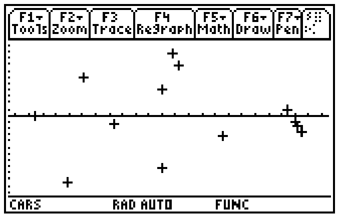
1. Select **y2** in [◆] [Y=].
2. Deselect **Plot 2**.
3. Define **Plot 1** as above, but with Mark: **Box** and Y List: **tmpg**.
4. Press [F5] **ZoomData** (screen 38).

   The residuals about the transformed linear regression line are saved in **statvars\resid**.

(38)


5. From the Stats/List Editor, turn off all functions and plots with [F2] **Plots**, **4:FnOff** and [F2] **Plots**, **3:PlotsOff**.
6. Define **Plot 2** as Plot Type: **Scatter**, Mark: **Plus**, X List: **wt**, Y List: **statvars\resid** (pasted from [2nd][VAR-LINK]).
7. Press [F5] **ZoomData** (screen 39).

   This is reasonably random.

(39)


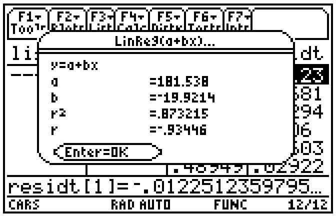
ln(x) Linear Transformation:

Transforming x can also straighten out data points.

From the Home screen, type: **ln(wt)→twt**

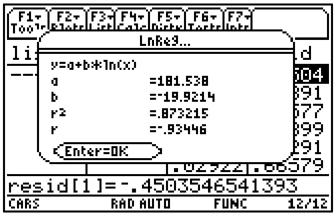
From the Stats/List Editor, press [F4]**Calc, 3:Regressions, 1:LinReg(a+bx)** on X List: **twt**, Y List: **mpg**, and Store to Reg: **y2(x)**.

Observe from screen 40 that r^2 = 87.3215%, mpg = 181.538 - 19.9214 * ln(wt) (with ***x*** replaced by **ln(wt)**) is the ln(x) linear transformation equation for **wt**.

(40) 

Instead of the process above, you can use the built-in **LnReg** option:

1. From the Stats/List Editor, press [F4] **Calc**, **3:Regressions**, **7:LnReg** with X List: **wt** and Y List: **mpg** can be executed to obtain the r^2 and regression equation as above and in screen 41 (with *x* replaced by **wt**).

(41) 

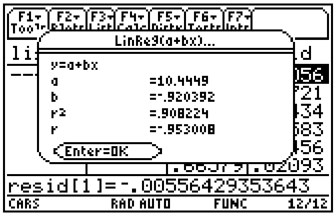
ln(x) and ln(y) Linear Transformation or Power Reg:

From the Home screen, type: **ln(wt)→twt**, **ln(mpg)→tmpg**.

From the Stats/List Editor, press [F4] **Calc**, **3:Regressions**, **1:LinReg(a+bx)** on X List: **twt**, Y List: **tmpg**, and Store to Reg: **y2(x)**.

Observe from screen 42 that r^2 = 90.8%, and the ln transformation equation is:

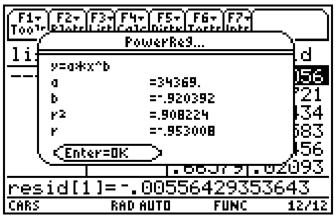
ln(mpg) = 10.4449 - 0.920392 ∗ ln(wt) (with ***y*** replaced by **ln(mpg)** and ***x*** replaced by **ln(wt)**. Exponentiating both sides of the equation, you obtain mpg = $e^{10.4449-0.920392\ \ln(wt)}$).

(42) 

Another option, using the built-in **PowerReg**:

1. From the Stats/List Editor, press [F4] **Calc**, **3:Regressions**, **9:PowerReg**, with X List: **wt** and Y List: **mpg** can be executed to obtain **mpg = 34369 * wt $^{-.920392}$** (screen 43).

2. Take **ln** of both sides for ln(mpg) = ln(34369 * wt $^{-.920392}$) = ln(34369) + ln (wt $^{-.920392}$) = 10.4449 - 0.920392 ∗ ln(wt), which is the same equation obtained above in the previous option.

(43) 

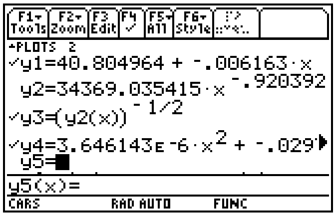
Summary: In looking at the first five transformations, you noticed r^2 increased from **83.8%** to **93.9%**.

The built-in commands of **ExpReg**, **LnReg**, and **PowerReg** save a lot of work, but it is helpful to look at them as linear transformations, to understand r^2, list **resid**, and list **residt**.

car	cyl	mpg	wt	Y1 Linear (|residual|)	Y3 Y^{-2} (|residual|)	Y4 Quadratic (|residual|)	* Smallest |residual|
Hyundai (Accent)	4	28	2290	26.7 (1.3)	28 (0)*	28.8 (0.8)	0
Mercury (Mystique)	6	20	3115	21.6 (1.6)	20.6 (0.6)	20.4 (0.4)*	0.4
Cadillac (Deville)	8	17	4020	16.0 (1.0)	16.8 (0.2)	17.0 (0.0)*	0
Mazda (Protégé)	4	29	2500	25.4 (3.6)	25.4 (3.6)	26.2 (2.8)*	2.8
Honda (Accord)	4	23	3245	20.8 (2.2)*	19.9 (3.1)	19.6 (3.4)	2.2

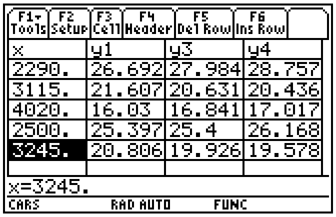
These predicted values are easily calculated by plugging into the appropriate equation and using Table Setup in **ASK** mode.

1. Press ◆ [Y=] and make sure that **y1**, **y3**, and **y4** are selected (screen 44).
2. Press ◆ [TblSet] and select Independent: **ASK**.
3. Press ◆ [TABLE].
4. Under the **x** column, type the weights of the five cars shown above.

(44) 

5. The values in the **y1**, **y3**, and **y4** columns will be the linear, y^{-2}, and quadratic predicted **mpg** values, respectively (screen 45).

The y^{-2} and Quadratic models work very well for all but the Mazda Protégé and Honda Accord, both 4-cylinder cars. In Topic 42 (Multiple Linear Regression), you will include other variables that better predict these two.

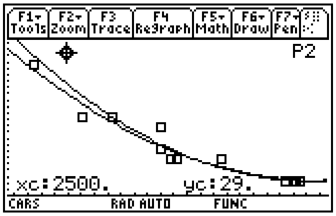
(45) 

x	y1	y3	y4
2290.	26.692	27.984	28.757
3115.	21.607	20.631	20.436
4020.	16.03	16.841	17.017
2500.	25.397	25.4	26.168
3245.	20.806	19.926	19.578

Note: *These are point estimates. Confidence Intervals are a more meaningful way to estimate and will be covered in Topics 33 and 42.*

Influential Observation

By removing the Mazda Protégé from the data set, r^2 for the Quadratic Regression changed from 91.1% to 94.4%. The curve shifts down for lower weights, but is nearly identical for cars that weigh more (screen 46). This data is also indicated in Topic 11, screen 13.

(46) 

The work in this chapter was done as an exploratory analysis. Sometimes, the physics of the problem or economic theory suggest the proper transformation. For example, the intensity of a light (***y***) is inversely proportional to the square of the distance from the light (***x***) which would suggest a x^{-2} transformation. If some variable (***y***) is increasing or decreasing at a constant percentage with time (***x***), mathematics theory can be used to prove **ln(y)** is the appropriate transformation, since the underlying model is exponential.

Other regression fits found in F4 **Calc**, **3:Regressions** will be covered in Topics 48, 49, and 50.