

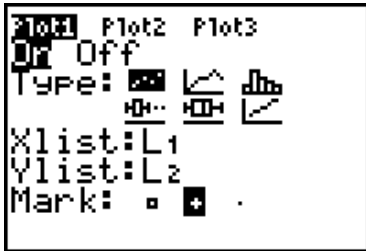
BIVARIATE DATA- LINEAR REGRESSION AP REVIEW SHEET

This unit has dealt with scatterplots of bivariate data and using the Least-Squares Regression Line (LSRL) to: ‘predict’ overall patterns; look for any deviations in the patterns noted; and give a way to add numerical descriptions of specific aspects of the data. While we have studied linear regression in previous courses, here we gain an understanding of the statistical significance, if any, that we may be able to determine from our regression models. Using simple examples, this review sheet will walk you through the unit and point out key features. **Be sure you enter all data and check your work against the diagrams provided!**

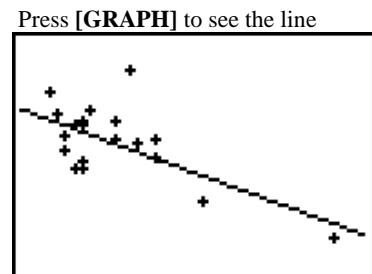
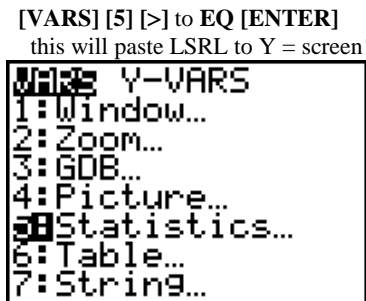
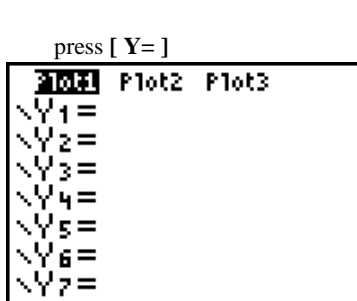
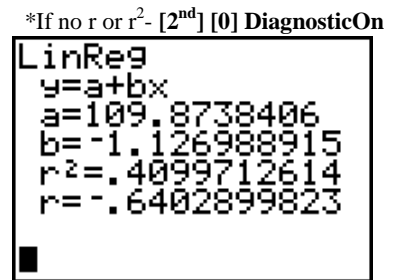
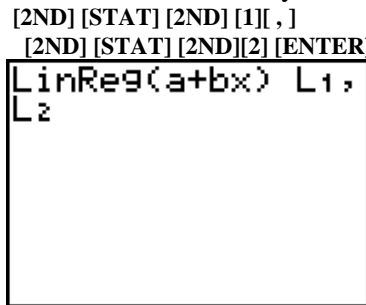
Ex. 1- Age (in months) at first word and the Gesell Adaptive Scores-(i.e .later mental ability)

Child	Age	Score	Child	Age	Score
1	15	95	12	9	96
2	26	71	13	10	83
3	10	83	14	11	84
4	9	91	15	11	102
5	15	102	16	10	100
6	20	87	17	12	105
7	18	93	18	42	57
8	11	100	19	17	121
9	8	104	20	11	86
10	20	94	21	10	100
11	7	113			

- Given data sets of bi-variate date (numerical, not categorical), create a scatterplot. and look for any notable patterns, etc. Adaptive score vs. Age ...i.e. age= x-axis, adaptive score = y-axis {in L₁ & L₂}



- Next, since this is bi-variate data, plot the LSRL. The key screens and strokes



Remember the actual LSRL is calculated as follows- $\hat{y} = a + bx$ where $\begin{cases} \text{slope is } & b = r \frac{s_y}{s_x} \\ \text{intercept is } & a = \bar{y} - b\bar{x} \end{cases}$

Next, we want to concentrate on how well this line actually ‘predicts’ what may occur given other children that were not measured. i.e. what about a baby who speaks after 5 or 6 months, at 15 ½ months, or not until they are 24 months old? Now, we are particularly interested in both the **correlation** and the **coefficient of determination**. (It should be noted, that while the formulas for each are provided and class demos for each were calculated, you will not be expected to hand-calculate these measures)

- **Correlation (r):** r tells just how well the two sets of data are associated. Do not confuse association with causation. Just because r shows a strong association, the 2 variables may be totally unrelated in terms of cause & effect. Correlation measures the strength and direction of the linear relationship between the two quantitative variables. **Our example reflects a reasonable negative association** (r = -.64). r is calculated by-

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

- **Coefficient of Determination (r²):** r² is the proportion of the total sample variability that is explained by the least-squares regression of y on x. You could say x becomes more of a poor predictor of y as SSM and the SSE become nearly the same numbers. This happens because as the numerator gets closer to 0, the entire proportion gets closer to 0%. **In this case, only 41% of the variation in y is explained by the LSRL of y on x.** (r² = .409)

SSM- Sum of the Square Means (also called TSS- Total Sum of Squares)

SSE- Sum of the Squares for Errors.

r² is calculated as follows-

$$r^2 = \frac{SSM - SSE}{SSM}$$

where $\begin{cases} SSM = \sum (y - \bar{y})^2 \\ SSE = \sum (y - \hat{y})^2 \end{cases}$

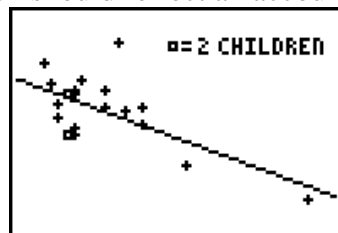
y = actual y data value

\bar{y} = mean of y data

\hat{y} = predicted y value from LSRL

NOW-- what do you see?

Look closely at the graph shown. The graph above is a little deceptive because there are two children that have exactly the same score. A sketch should reflect an accounting for this by using a different symbol, etc.



In the picture above, two notable things stand out. Child 18 on the bottom right and child 19 near the top of the screen. In a moment we will talk of points like these as being either an **outlier** or **influential** once we will look at the residuals.

Finally we want to study the fit of the regression line and draw conclusions, if possible, for future use/ study. This requires that we also look at the **residuals** and **residual plots**. In addition, be mindful of points that have unusually large residual, nonlinear patterns, and uneven variation about the residual plot line.

- **Residuals**-the difference between an observed value of the response variable and the value predicted by the regression line. $residual = observed\ y - predicted\ y$ ($y - \hat{y}$) If you calculate the mean of the residuals you will always get a mean of zero.
- **Residual Plots**- Since the residual is $y - \hat{y}$, each residual is either a positive distance or a negative distance from the LSRL line. If the residuals were plotted against the horizontal line $y = 0$, you could easily see the positive and negative residual distances. The reason you would want to plot the residuals in the first place is to note any unusually large residuals, watch for nonlinear patterns, and uneven variation about the residual plot line. To do this on your calculator, you can use the following procedure.
 1. Enter your scatterplot and perform the linear regression as shown earlier.
 2. Insert a new list named **RES** beside your original lists of **L1** and **L2**. #1 below.
 3. Highlight **RES** and define this list as the observed value minus the predicted value (**L2- Y1(L1)**). #1 & #2 below (Note: this **RES** you just created will have the EXACT same values as those found in **RESID** list that your calculator creates **after** you run the LSRL.)
 4. Next, disable **Y1**, the regression equation and make **Y1=0**. #3
 5. Finally, set Plot2 with **L1** as the *x*-variable and **RES** as the *y*-variable.#4below.
 6. **ZOOM 9** will graph the residual plot. #5 below.

#1: enter argument for RES

L1	L2	RES	3
15	95	-----	
26	71		
9	91		
15	102		
20	87		
18	93		
11	100		
RES = L2 - Y1(L1)			

#2: press [ENTER] to see calculation

L1	L2	RES	3
15	95	2.031	
26	71	-9.572	
9	91	-8.731	
15	102	9.031	
20	87	-3.341	
18	93	3.412	
11	100	2.523	
RES(1) = 2.03099313...			

#3: press [Y=] [=] to Y2, enter [0]

Plot1	Plot3
Y1=109.87384058	
517+-1.126988914	
8631X	
Y2=0	
Y3=	
Y4=	
Y5=	

#4: [2ND] [Y=] [2] & select as shown

Plot1	Plot3
Off	
Type: [] [] [] []	
Xlist: L1	
Ylist: RES	
Mark: [] + []	

#5: [ZOOM] [9]



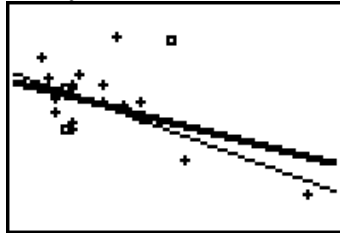
Based on the above plot, now you can begin to study the effect of the residuals and look for ‘peculiarities’.

First, look for outliers or influential points.

- **Outlier**- an observation that lies outside the overall pattern of the other observations in a scatterplot. An observation can be an outlier in the *x* direction, the *y* direction, or in both directions.
- **Influential**- an observation is considered to be ‘influential’ if removing it would markedly change the position of the regression line. Points that are outliers in the *x* direction are often influential. Influential

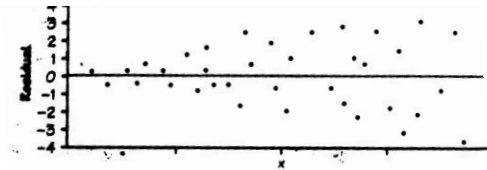
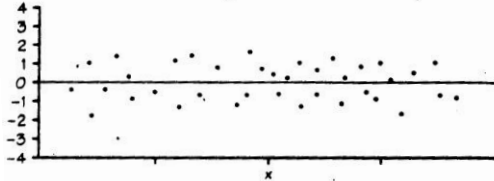
points often have small residuals because they pull the regression toward themselves. *See picture below.* Looking back to the original graph, notice the point on the bottom right (child 18) and point near the top (child 19). Both seem to ‘pull away from’ the rest of the data. In this case, we would say that Child 18 is an outlier in the x direction and is influential as seen in the following picture. Child 19 is an outlier in the y direction.

Recall the original graph from page 1,2
Dark line reflects LSRL without child 18

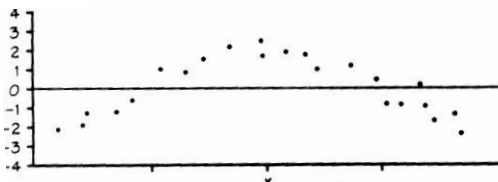


Second, look for a linear or nonlinear pattern in the actual residual plot.

- **Ideal residual plot-** Notice the residuals are all plotted closely to the $Y=0$ line and in a uniform, nice flowing pattern.
- **Increasing/ decreasing spread for larger values-** The response variable y in the plot has more spread for larger values of the explanatory variable x , so prediction will be less accurate as x gets large.



- **Nonlinear-** The data has a curved pattern, so a straight line is an inappropriate model for the data.
- **Watch for large residuals Or points extreme in x direction-** Large residuals, again like Child 19, are outliers because they lie outside the straight-line pattern. In an extreme x -direction, like Child 18, may have small residuals, but are still very influential.



Finally, once you have completed all this you can make a final analysis of the patterns, trends, conclusions that may be drawn from your data sets. How do you decide what’s important? Before you can write, you must first consider everything you have seen and what it all means.

- **In the Score vs. Age data example-** You will note that the original data have $r^2 = 0.41$, That is, the age at which a child begins to talk explains 41% of the variation on a later test of mental ability. This relationship is strong enough to be interesting to parents. But if you leave out child 18, r^2 drops to only 11%. What should the child development researcher do? She must decide whether Child 18 is so slow to speak that this individual should not be allowed to influence the analysis. If she excludes Child 18, most of the evidence for the connection between the age at which a child begins to talk and later ability level vanishes. If she keeps Child 18, she needs data on other children who were also slow to begin talking, so that the analysis no longer depends so heavily on just one child. Careful consideration needs to be given to each item you choose as important in your final analysis!

REVIEW PROBLEMS-

- 1- Manatees are large, gentle sea creatures that live along the Florida coast. Many manatees are killed or injured by powerboats. Here are data on powerboat registrations (in thousands) and the number of manatees killed by boats in Florida in the years 1977 to 1994.

Year	Boat registr- (100)	Manatees Killed	Year	Boat registr- (100)	Manatees Killed
1977	447	13	1986	614	33
1978	460	21	1987	645	39
1979	481	24	1988	675	43
1980	498	16	1989	711	50
1981	513	24	1990	719	47
1982	512	20	1991	716	53
1983	526	15	1992	716	38
1984	559	34	1993	716	35
1985	585	33	1994	735	49

- We want to examine the relationship between number of powerboats and number of manatees killed by boats. Which is the explanatory (or x) variable?
 - Make a scatterplot of these data. Be sure to label the axes with the variable names, not just x and y . What does the scatterplot show about the relationship between these variables/
 - From your scatterplot, does it appear that there is a strong straight-line pattern? What is r^2 for these data?
 - Draw the regression line on your scatterplot. Predict how many manatees would be killed each year if Florida decided to freeze the number of boats at 700,000.
 - Circle on your plot the one observation that has a somewhat large residual. (we have no reason to remove it.)
 - This particular distribution is roughly normal. Using the 68-95-99.7 rule for normal distributions, say how surprising a residual with standardized value -2.08 is.
- 2- The mean height of American women in their early twenties is about 64.5 inches and the standard deviation is about 2.5 inches. The mean height of men the same age is about 68.5 inches, with standard deviation about 2.7 inches. If the correlation between the heights of husbands and wives is about $r = 0.5$, what is the slope of the regression line of the husband's height on the wife's height in young couples? Draw a graph of this regression line. Predict the height of the husband of a woman who is 67 inches tall.

See the attached additional problems to prepare for your test as well.

Be sure to review the attached correlation examples as well for correlation 'estimation' practice.