AP Statistics
Unit 02 – Bivariate Data
Day 03 Notes: Residuals and Influential Points

Name Key
Period

Residual: the difference between an observed value of the response variable and the value predicted by the regression line.

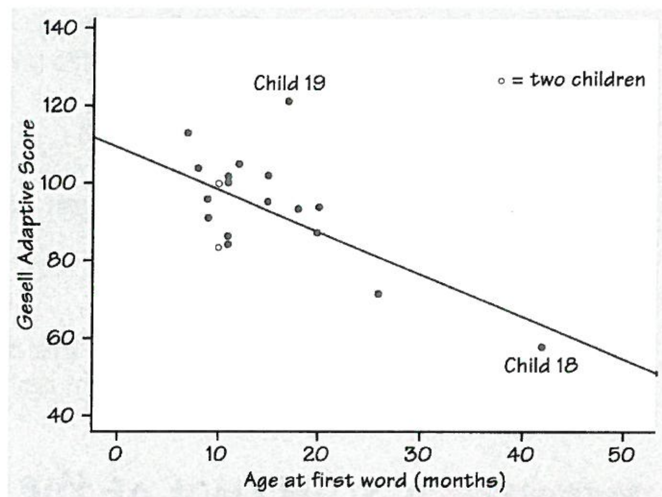
Residual= observed y – predicted y
Residual= y - $\hat{y}$

Example:

TABLE 3.4 Age at first word and Gesell score

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

Source: These data were originally collected by L. M. Linde of UCLA but were first published by M. R. Mickey, O. J. Dunn, and V. Clark, "Note on the use of stepwise regression in detecting outliers," *Computers and Biomedical Research*, 1 (1967), pp. 105–111. The data have been used by several authors. We found them in N. R. Draper and J. A. John, "Influential observations and outliers in regression," *Technometrics*, 23 (1981), pp. 21–26.



LSRL by hand: $\widehat{\text{Gessell score}} = 109.87 - 1.1270(\text{age at first word})$

$\bar{x} =$ 14.3810 $\quad S_x = 7.9465$

$\bar{y} =$ 93.6667 $\quad S_y = 13.9869$

r = −0.6403

b = $r\frac{S_y}{S_x} = -0.6403\left(\frac{13.9869}{7.9465}\right) = \boxed{-1.1270}$

a = $\bar{y} - b\bar{x} = 93.66667 + 1.1270(14.3810) = \boxed{109.87}$

R^2 = (in context) 41.00% of the variance in Gessell score can be accounted for by the LSRL of Gessell score on age at first word.

To calculate a residual:

- Put in data in L1 and L2
- Get a line of regression (8. LinReg a+bx)
- Once you have the regression formula, put in a data point for x
- That will give you the predicted $\hat{y}$
- Subtract your predicted y from the observed y

example:

$\widehat{GS} = 109.87 - 1.1270(age)$

$x = 26$ $\quad \widehat{GS} = 109.87 - 1.1270(26)$

$y = 71$ $\quad \widehat{GS} = 80.5721$

residual $= GS - \widehat{GS} = 71 - 80.5721$

$= -9.5721$

To calculate all residuals and put them in a list (to graph, if you'd like):

- Put in data in L1 and L2
- Get a line of regression (8. LinReg a+bx) – this make your calculator automatically calculate residuals and places them in the RESID list!
- Go back to your lists and highlight L3
- Click 2nd LIST and choose 7: RESID, then click enter
- Your residuals now appear in L3

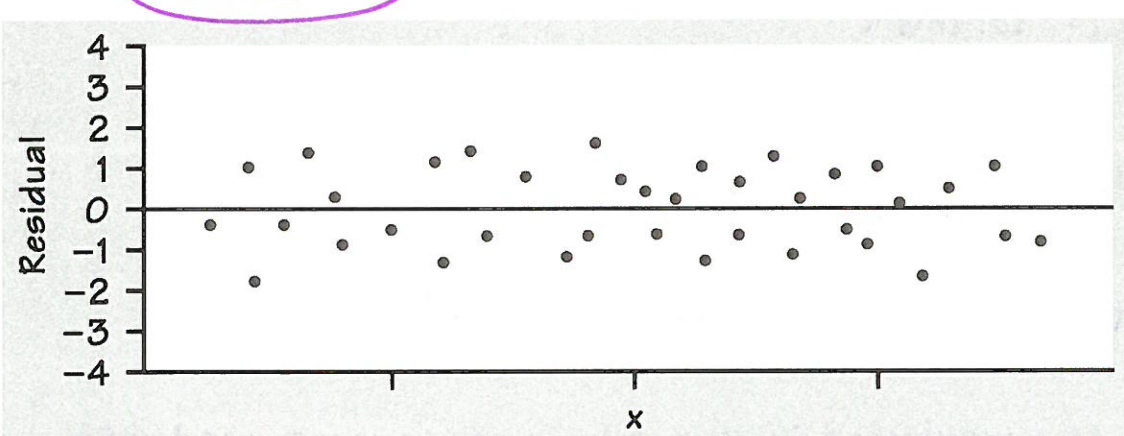
There is a residual for each data point. **The mean of the least-squares residuals is always ZERO.** When you do it in a calculator you might not get exactly zero because of their rounding. This is called a **"roundoff" error."**

Residual Plot: a scatterplot of the regression residuals against the explanatory variable.

- helps us assess the fit of a regression line!

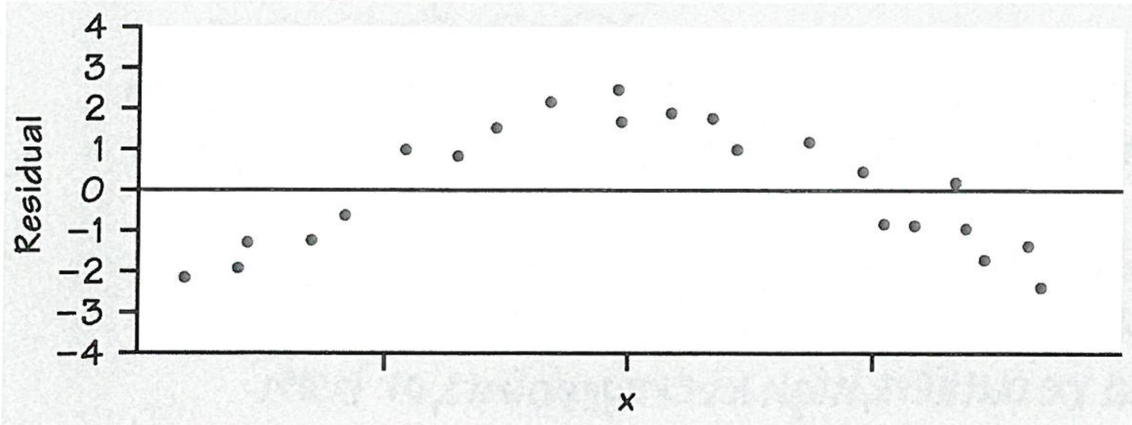
What to look for while examining residuals:
The **uniform** scatter of points indicates that the regression line fits the data well, so the line is a good model.

random

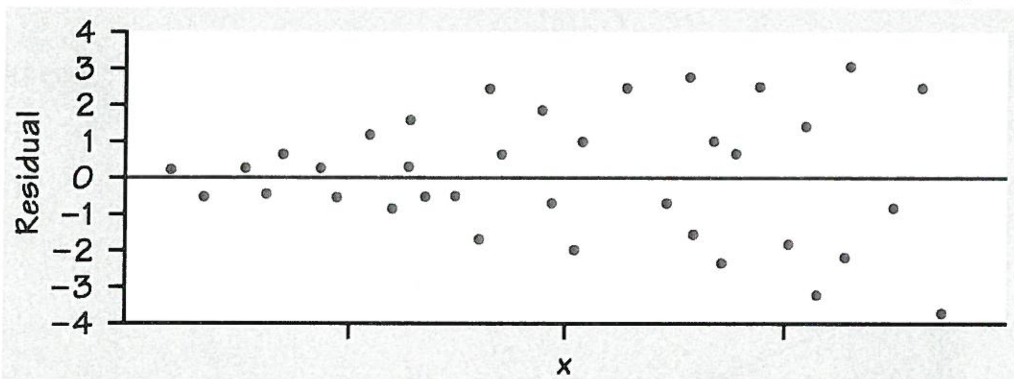


DON'T WANT

1. The residuals have a **curved** pattern, so the straight line is an inappropriate model.



2. The response variable(y-variable) has more spread for the larger values of the explanatory variable (x-variable) so prediction will be less accurate when x is large.



3. Individual points with large residuals like outliers in the vertical (y) direction because they lie far from the line that describes the overall pattern

4. Individual points that are extreme in the x direction with smaller residuals, but they can be very important.

Outlier: an observation that lies outside the overall pattern of the other observations. * in the y direction

High leverage: an observation that lies outside the overall pattern of the other observations. * in the x direction

Influential observations: a data point that, if removed, would change the result of the line (LSRL equation)

* could be outliers, high leverage points, or both

- Influential points often have small residuals because they pull the regression line towards themselves
- Influential observations can greatly can greatly change the interpretation of the data

Solid line is calculated from all the data. The dashed line is calculated leaving out Child 18. Child 18 is an influential observation because leaving out this point moves the regression line quite a bit.