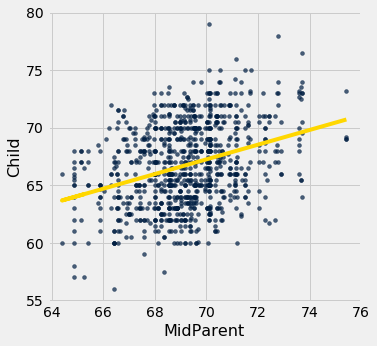
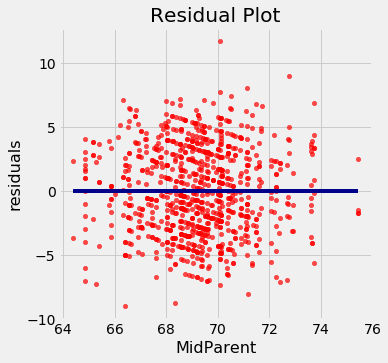
**Data 8 Spring 2020**

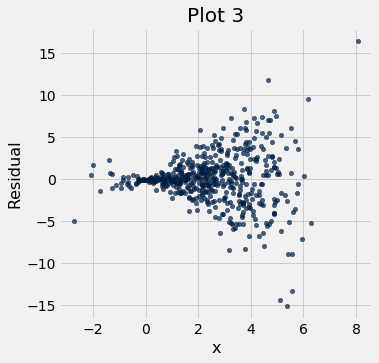
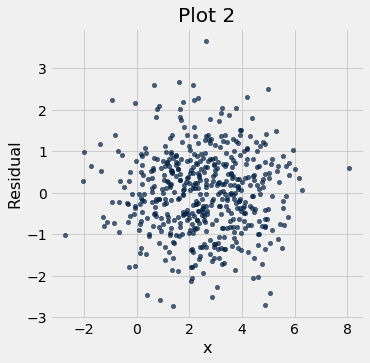
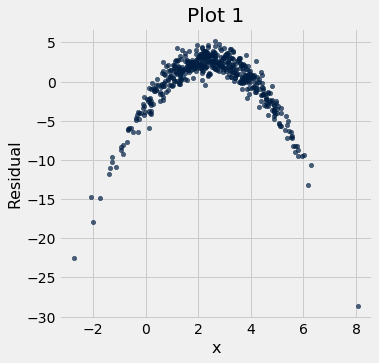
**Discussion: Residuals and Regression Inference**

In data science, we can use regression inference in order to make predictions; however, in order to assess the accuracy of our linear regression model, we want to examine the error between our predictions and the actual data. These errors are called *residuals*.

An example can be found below in the graph of midparent heights compared to child heights. The graph of the residuals is shown on the right.

** **

**Question 1.** Displayed below are three residual plots. For which of the following residual plots is using linear regression a reasonable idea, and why?

****

Plot 2 and Plot 3 are residual plots for which using linear regression on the original data is reasonable. Plot 2 is the best residual to use linear regression for, since the residuals have the pattern of a formless cloud. Note that using linear regression for plot 3 is reasonable, but the residual plot is heteroscedastic, which means that the residuals are unevenly spread for different x values (check out the textbook for examples!)

**Question 2.** Yash has a sample of 100 snacks (Yum!). This dataset contains the calories from fat (cal\_fat) and the calories total (cal\_total) for each snack. Yash wants to use a snack’s cal\_fat to predict its cal\_total. The standard deviation of cal\_fat is 5 calories, and the standard deviation of cal\_total is 10 calories. The correlation coefficient between the two variables is 0.6.

a. What would be the SD of the residuals between the predictedcal\_total and the actual cal\_total?

**(If r = 0.6)**

SD of residual = ((1 - r\*\*2)\*\*½ )\*SD of y

SD of residual = 10\*(0.64)\*\*½

SD of residual = 10\*0.8 = 8

b. Suppose the correlation coefficient between the two variables was actually 0.9. What would be the SD of the residuals?

**(If r = 0.9)**

SD of residual = ((1 - r\*\*2)\*\*½ )\*SD of y

SD of residual = (1 - 0.81)\*\*½ \* 10

SD of residual = (.19)\*\*½ \* 10

SD of residual = .44\*10 = 4.4

c. What does this say about the relationship between the SD of the residuals and the correlation coefficient?

This says that as the correlation goes up, the SD of the residuals goes down, which makes sense since the residuals are errors and if they are more correlated, then the errors in our predictions will go down.

d. Yash thinks that there is no association between cal\_fat and cal\_total, and that his sample was just biased. How can Yash test this hypothesis?

Null Hypothesis: There is no association between fat calories and total calories, the true correlation is 0.

Alternative Hypothesis: The association between fat calories and total calories is not zero.

Describe Testing Method: Yash should bootstrap his sample repeatedly, generate a confidence interval for the correlation and check to see if zero is in the CI.

e. Yash runs his hypothesis test and gets a 99% confidence interval of 0.24 to 089. Should he reject the null hypothesis?

Yes, Yash should reject the null hypothesis, because 0 is not in the CI.

f. Finally, Yash wants to generate a line of best fit for his data. Should he use the method of least squares or the regression equations?

It doesn’t matter what method Yash uses, they both return the same line!