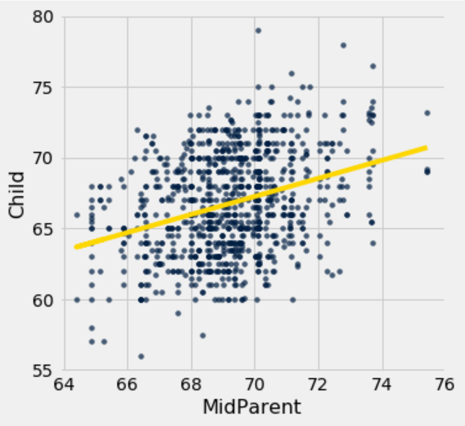
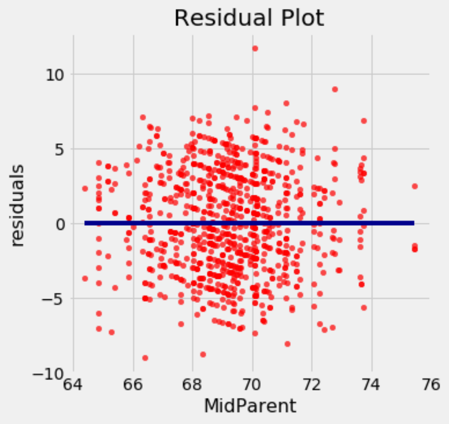
CS2065 Spring 2022
Discussion: Residuals and Regression Inference

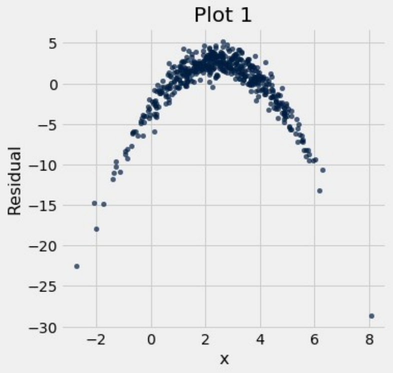
In data science, we can use regression inference in order to make predictions; however, in order to assess the accuracy of our linear regression model, we want to examine the error between our predictions and the actual data. These errors are called *residuals*.

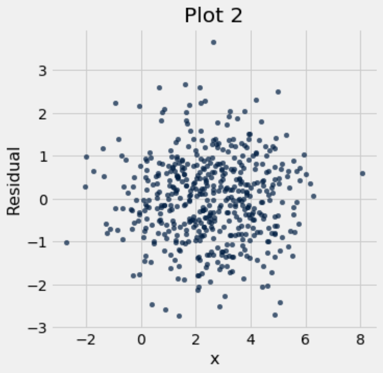
An example can be found below in the graph of midparent heights compared to child heights. The graph of the residuals is shown on the right.

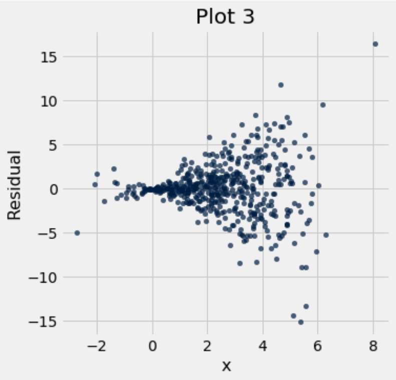




Question 1. Displayed below are three residual plots. For which of the following residual plots is using linear regression a reasonable idea, and why?







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 predicted `cal_total` and the actual `cal_total`?

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

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

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:

Alternative Hypothesis:

Describe Testing Method:

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

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?