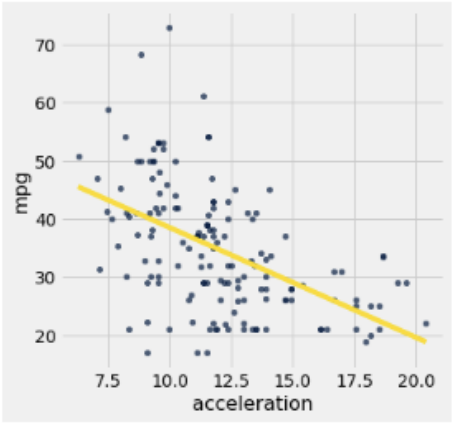
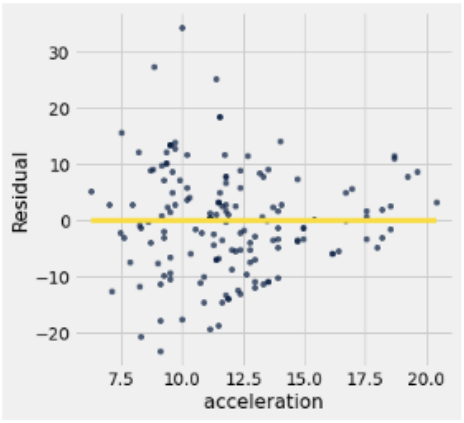
In data science, we can use linear regression in order to make predictions. Moreover, we want to assess the accuracy of our predictions. To do so, we can examine the error between our actual data and the predictions; these errors are called *residuals*.

An example can be found below in the graph of miles per gallon compared to acceleration. The graph of the residuals is shown on the right. The yellow line is our regression line.



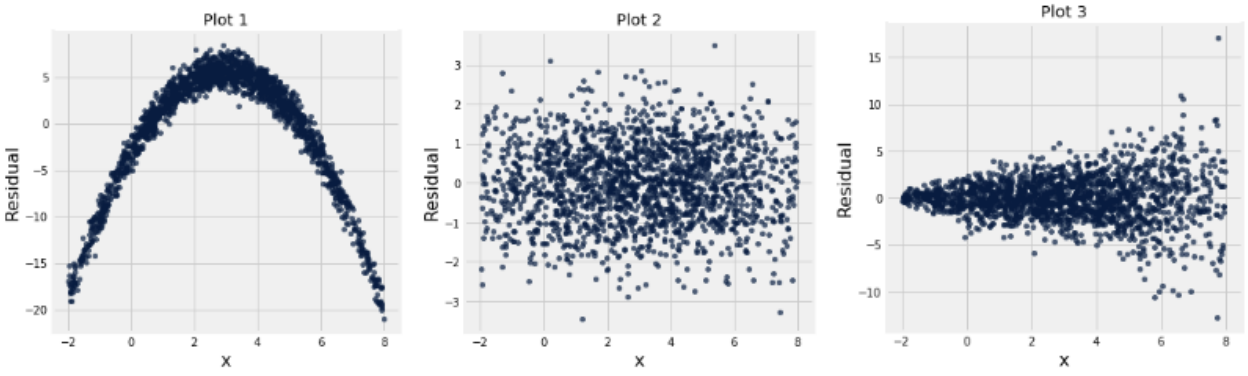


As a reminder:

- residual = y - estimated value of y = y - height of regression line at x
- The mean of residuals is zero and they show no trend (i.e. correlation is zero)

1. Visual Diagnostic

Displayed below are three residual plots. For which of the following residual plots is using linear regression a reasonable idea, and why? What might the original graphs have looked like?



2. Fight for California!

At a Cal Football Game the Mic Men, spirit leaders, claimed that our opponent's ability to score was linearly affected by the student section's noise level. Atticus thinks they're wrong. Instead he believes there is no linear relationship between student section noise and opponents scores. A friend gives Atticus a table `noise` which contains the following information:

- `Opponent`: Cal's opponent for the game
- `Loudness`: Each row represents the maximum noise in decibels produced by the Student Section when Cal is on defense during a single home football game. *These are fabricated by Atticus's friend
- `Points`: The number of points scored by the opposing team.

A sample of the table with data from the 2023 season is below.

Opponent	Loudness	Points
California-Davis	65	13
USC	55	50
Arizona State	75	21
2 Washington	85	59
Oregon State	70	52
Stanford	120	15
UCLA	100	7

a) Atticus thinks that there is no correlation between `Loudness` and `Points`, and that the Mic Men's claim is wrong. How can Atticus test his hypothesis?

Null Hypothesis:

Alternative Hypothesis:

Describe Testing Method:

b) Atticus decides to write a function which produces one bootstrapped estimate of the correlation between `Loudness` and `Points`. Define the `one_relationship` function below which takes in the following arguments:

- **table** (Table): a table with the data points like `noise`
- **x** (string): the column name for the x variable
- **y** (string): the column name for the y variable

The function should return a float by using the sample data to produce one bootstrapped estimate of the correlation. You can assume the function `correlation(x,y)` returns the correlation between arrays `x` and `y`.

```
def one_relationship(table, x, y):
    bootstrap = _____
    x_values = _____
    y_values = _____
    return _____
```

For convenience here's another image of the sample of the table.

Opponent	Loudness	Points
California-Davis	65	13
USC	55	50
Arizona State	75	21
2 Washington	85	59
Oregon State	70	52
Stanford	120	15
UCLA	100	7

c) Atticus decides to generate a 70% confidence interval for the true correlation between `Loudness` and `Points` using 1000 bootstraps. Fill in the following code to generate the interval.

```
correlations = _____________________________________

for i in _____________________:

     bootstrapped_correlation = _____________________________________

     correlations = _____________________________________

left = _____________________________________

right = _____________________________________
```

d) Atticus enjoys chaos so he decides to swap the x and y arguments each time he makes a call to one_relationship inside his for loop. Should this impact his interval?

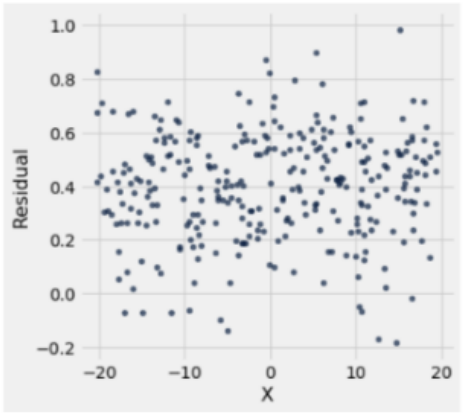
e) After running the above code Atticus gets an interval of [-0.75, -0.14]. Can the Mic Men claim Atticus is wrong and there is actually a direct causal effect between crowd noise and opposing team performance?

f) Regardless, Cal Athletics wants you to generate a line of best fit for your data. Should you use the method of least squares (i.e. minimizing RMSE) or the regression equations? Is there a difference between the two?

3. Sp19 Final Q9

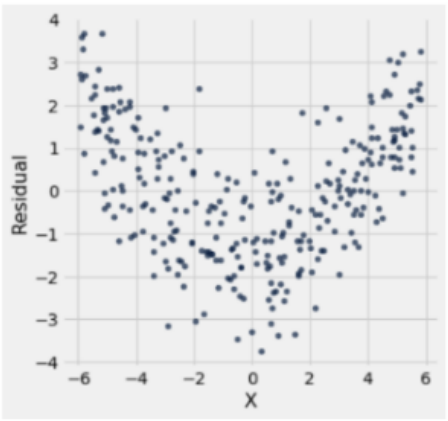
Each of the following plots represents the residuals from an attempted linear regression of a variable Y on a variable X (that is, the regression line is meant to predict values of Y based on values of X). For each one, indicate whether the regression line seems to be a good fit, or seems to be a bad fit, or if it is impossible for a residual plot to look like the plot shown. In each part, choose the best option based on what you see in the plot and just rough mental math if needed. Don't attempt any precise calculations.

(a) (2 pt)



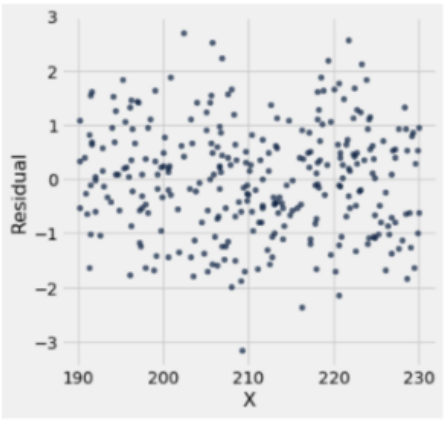
- ○ The regression seems to fit the data well
- ○ The regression seems not to fit the data
- ○ A residual plot could never look like this

(b) (2 pt)



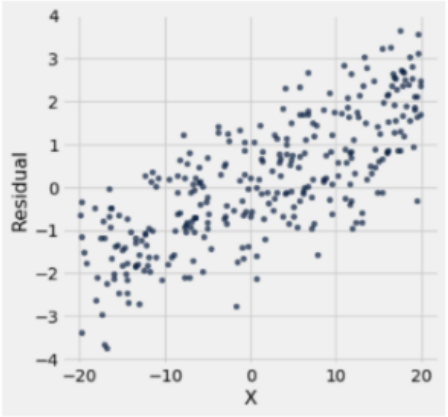
- ○ The regression seems to fit the data well
- ○ The regression seems not to fit the data
- ○ A residual plot could never look like this

(c) (2 pt)



- ○ The regression seems to fit the data well
- ○ The regression seems not to fit the data
- ○ A residual plot could never look like this

(d) (2 pt)



- ○ The regression seems to fit the data well
- ○ The regression seems not to fit the data
- ○ A residual plot could never look like this