Data 8 Tutoring – Week 11

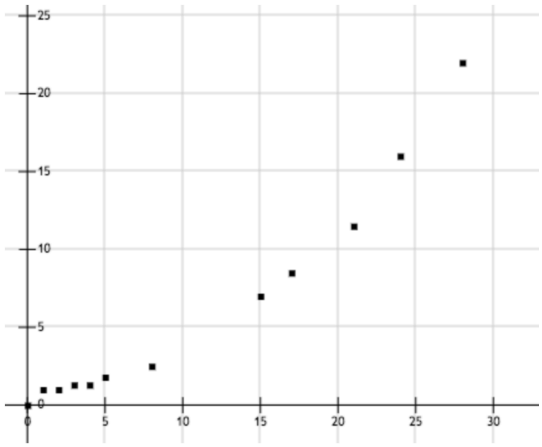
Regression

Recall:

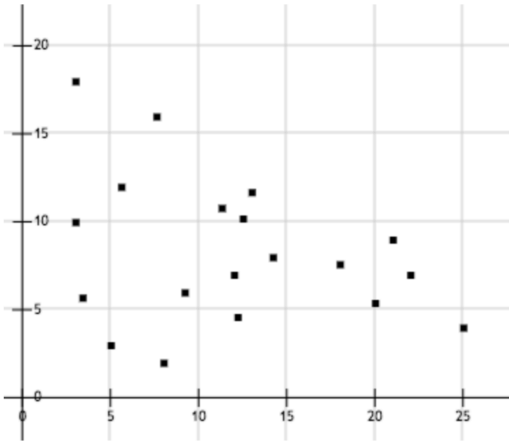
$$\text{residual} = \text{observed value} - \text{regression estimate}$$

The **residual** is the vertical distance between a data point and the residual line. Creating a residual plot can help you detect nonlinearity and heteroscedasticity in your data.

1. Draw the approximate regression line and residual plot for the following data. What conclusions can you come to from looking at the residuals?



Graph 1



Graph 2

2. How is the spread of the residuals related to the accuracy of our prediction? What equation describes this relationship?

3. Given a data set relating variables x, y **with correlation** $r = 0.45$ **and with variance of residuals equal to** 25**, what is the variance of the** y **values?**

Regression Inference

The Regression Model: The regression model states that our scatter plot of data resembles points generated by adding random noise to points originally along a straight line (the true line). We approximate this by finding the regression line. When applying regression inference techniques, we assume this model holds true for our data.

Regression Inference: What if the data we have is only a sample from a larger population? If we see a correlation in the data, is it due to random chance or is it representative of the population? And if we make a prediction using this line, how

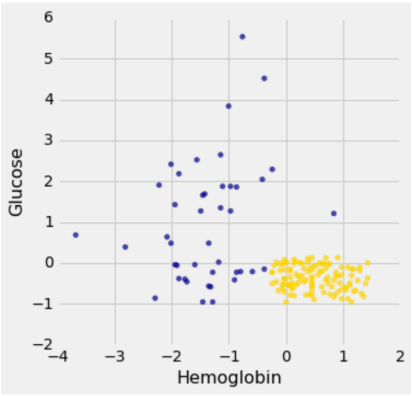
accurate will it be? We can apply the inference techniques we already know to regression problems as well.

4. Describe the process of testing to decide if the regression slope we are seeing is real, or just due to sampling variability.

Nearest Neighbors Classification

5. Given a data set from which we want to create a classifier, what groups should we split our data set into, and what are the purposes of each group?

6. Assume a 1-nearest neighbors classification is applied to the data set shown to the right, sketch roughly where the decision boundary would lie. (Taken from textbook)



7. Why do we ever use a value of k greater than 1 when classifying using the k-nearest-neighbors method? In other words, why do we look at more than just the single closest neighbor to a point to classify it?