Data 8 Tutoring – Week 11

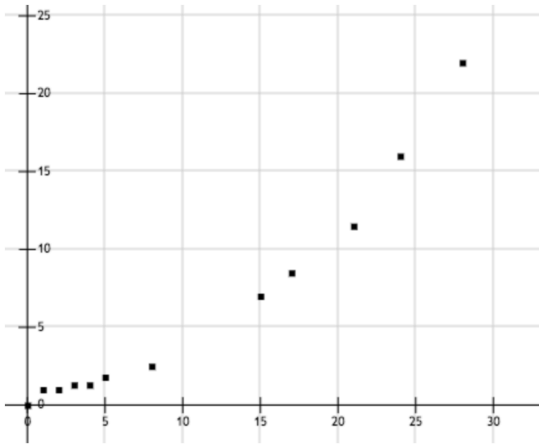
Regression

Recall:

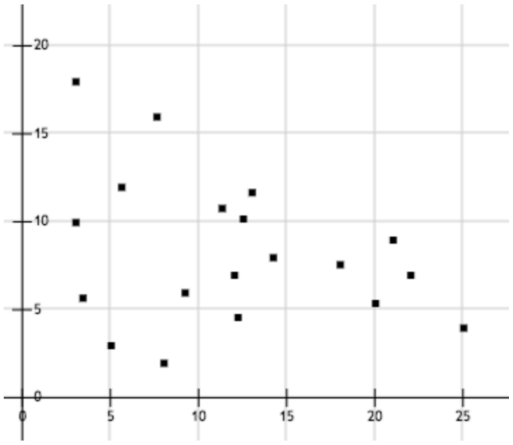
$$\text{residual} = \text{observed value} - \text{regression estimate}$$

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

Graph 1: There is a trend in the residual plot, with points further to the right more likely to have a positive residual. This suggests non-linearity in the data, and perhaps a quadratic best-fit line would be better.

Graph 2: The residuals on the left of the graph are larger in scale, while the ones on the right are smaller. Though the data seems to be linear, this pattern shows heteroscedasticity (uneven spread) in the data, and should be taken into account when making predictions (ie. some predictions may be more or less accurate than others).

Note to tutors: Let your students know that they don't have to be very accurate. All that's necessary is that they get the general shape of the plot down and can spot the trend; a rough sketch is sufficient.

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

$$SD_{residuals} = \sqrt{1 - r^2} \cdot SD_y$$

The standard deviation of the residuals is inversely related to r, which measures the correlation of our data. The stronger the correlation, the closer our residuals are to zero and the more accurate our prediction (remember, the standard deviations of the residuals are centered around 0).

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?**

$$SD_y = \frac{SD_{residuals}}{\sqrt{1 - r^2}}$$

$$Var_y = \frac{Var_{residuals}}{1 - r^2} = \frac{25}{1 - 0.45^2} = 31.34796$$

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.

Process:

- Describe the null hypothesis (true slope is 0) and alternate (true slope is not 0)
- Decide on the confidence interval and P value
- Use the bootstrap technique to generate random samples of data
- Calculate the regression slope for each sample
- Plot the slopes, generate the confidence interval
- See if slope = 0 is within this confidence interval, then decide to reject or fail to reject the null.

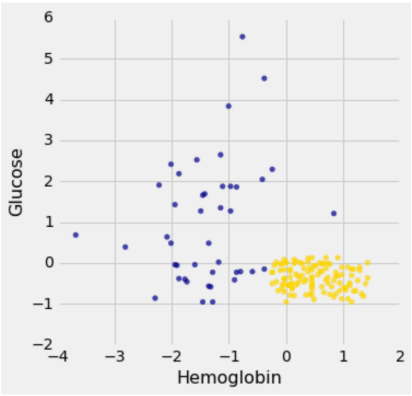
Nearest Neighbors Classification

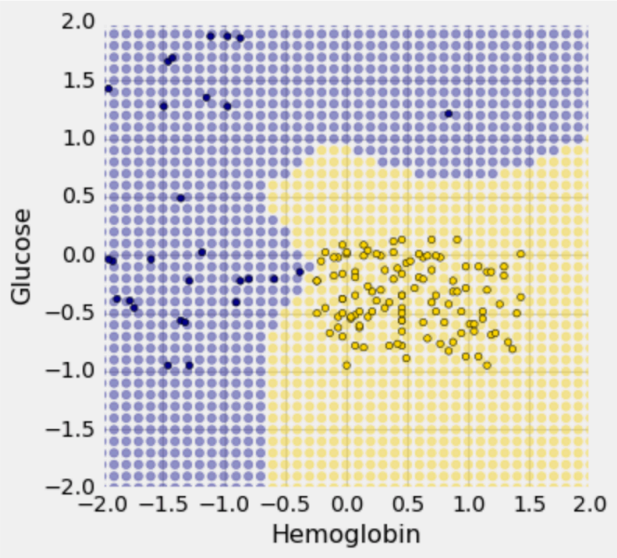
5. Given a data set from which we want to create a classifier, what groups should we split our data set into, and what are the purposes of each group?

The following groups:

- Training set – the data that we use to create our classifier
- Testing set – the data that we apply our classifier to in order to check its accuracy

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?

If we use more than just the closest point to our selected point to predict our point's classification, we are using more data and are likely to obtain a more accurate result. This is mainly true for points that are near boundaries. If the single closest point to a point P is yellow but the next four closest to it are blue, the first strategy would predict that P is yellow while the second would predict P to be blue.