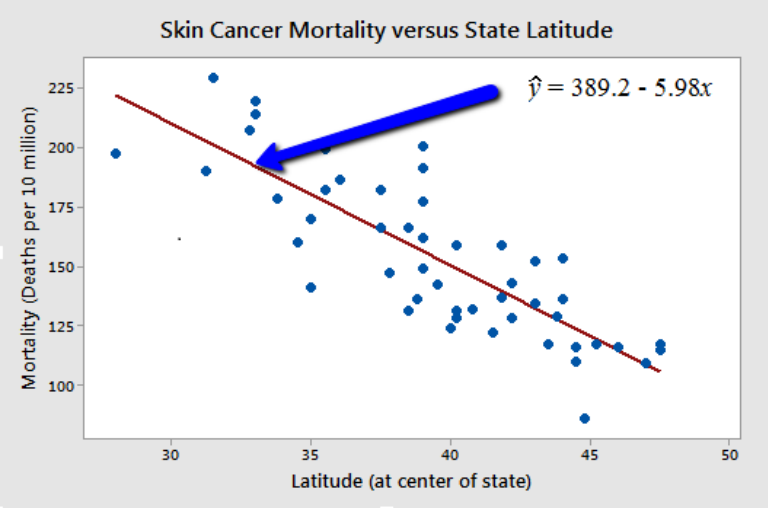
Exploring Linear Regression

Part #1: Regression Lines

Consider the scatterplot above, recording skin cancer mortality rates (# of deaths per 10 million people) in the 1950s against latitude for 48 states (Alaska and Hawaii were not states at the time of the data collection). Answer the following questions:



1. What is the explanatory variable? What is the response variable?

2. What kind of association do these variables exhibit? Explain.

3. What is the y-intercept of the regression line? Interpret this value in the context of the problem.

4. What is the slope of the regression line? Interpret this value in the context of the problem.

5. Pennsylvania is at a latitude of 40.8°. Using the regression line, how many skin cancer deaths would you predict for Pennsylvania residents?

6. The actual number of skin cancer deaths reported in Pennsylvania was 132. How does this compare to your predicted value? Compute the residual.

7. California is at a latitude of 37.5°. What value do you predict for California skin cancer deaths?

8. The actual observed value for California was 182. How does this compare to your predicted value? Compute the residual.

9. The "best-fit" regression line for a data set is chosen by finding the line that minimizes the sum of all of the residual values *squared*; that is, the chosen line minimizes

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

where (x_i, y_i) is the actual observed point on the scatterplot, $\hat{y}_i$ is the predicted value at x_i, and n is the total number of data points. We often refer to this method as least-squares regression. Why do you think we use the sum of *squares* of the residuals, and not just the straightforward sum of the residuals? (Hint: Think about your answers to #6 and #8 in this context).

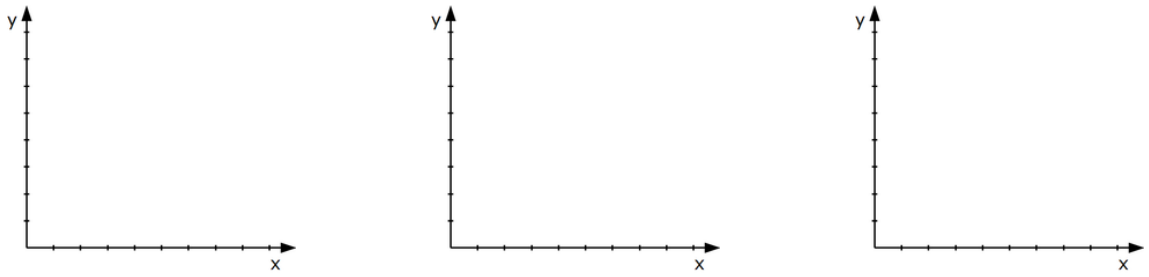
Exploring Linear Regression

Part #2: Correlation

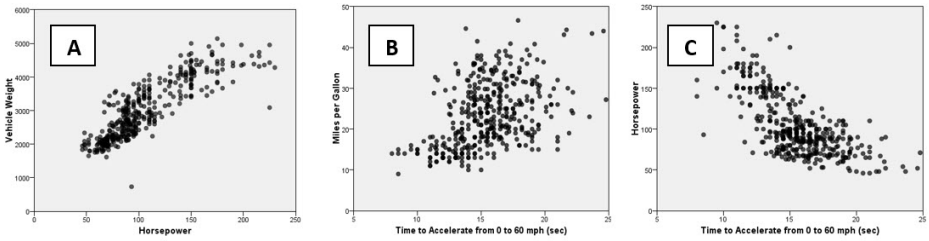
Correlation measures the strength of the linear relationship between two variables. We say two variables are strongly correlated if their scatterplot is clustered closely about the regression line, and weakly correlated if the association is looser.

1. Draw example scatterplots that depict the following correlation types:

   a) Strong positive correlation b) Weak negative correlation c) No correlation



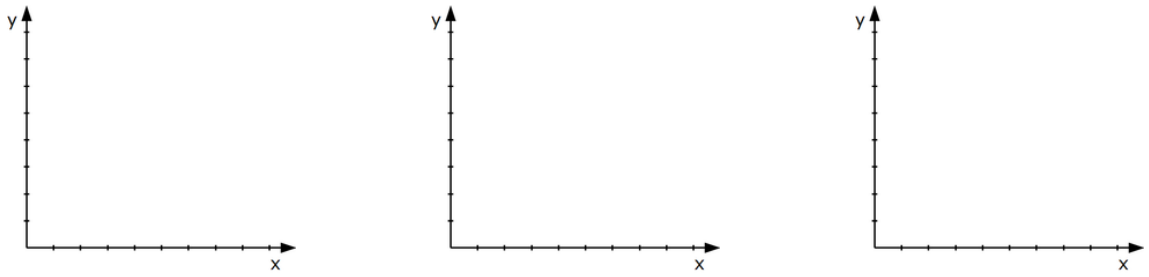
Correlation is always assigned a value between $r = -1$ and $r = 1$. The sign indicates whether the linear association is positive or negative, while the magnitude gives information about the strength of the relationship. A perfect positive association will have correlation $r = 1$. If two variables have no association at all, their correlation is $r = 0$. See below for examples:



Correlation =0.86 | Correlation =0.43 | Correlation = - 0.7

2. Draw example scatterplots that depict the following correlation values:

   a) $r = 0.5$ b) $r = -0.95$ c) $r = 0.05$



We often use r^2 (called the *squared correlation* or the *coefficient of determination*) to assess the strength of the relationship between the explanatory and response variables. The squared correlation lies in the interval $[0, 1]$ and describes how much of the variation in the data can be predicted by the regression line. The typical interpretation of the r^2 value is "the percent of variation in y that can be explained by x".

3. Suppose that height and weight are positively associated with a correlation coefficient of $r = 0.7$. What percentage of the variation in weight can be explained by the variation in height?

4. The relationship between temperature and ice cream sales can be characterized by $r^2 = 0.915$. Interpret this value in the context of the situation.

5. Suppose that number of hours spent watching TV per week and college GPA are characterized by a linear relationship with $r = -0.6$. What percentage of the variation in college GPA is *unexplained* by its relationship with number of hours spent watching TV?

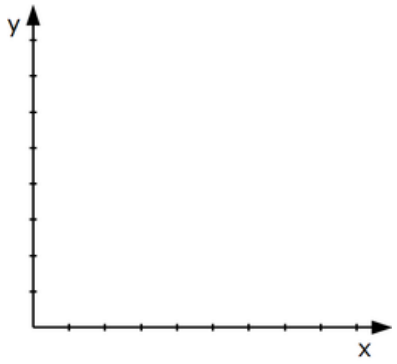
Exploring Linear Regression

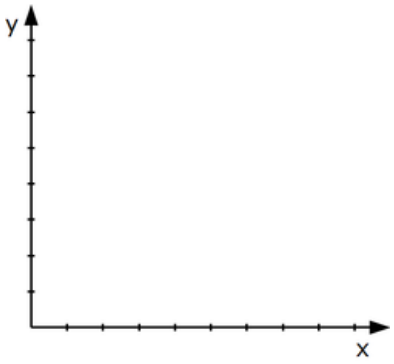
Part #3: Issues in Regression and Correlation

There are a number of issues that we need to be aware of when using regression lines and correlation to characterize our data. Work through the following questions to learn about just a few of them!

Outliers:
The presence of outliers can have a major impact on the equation of your regression line and the correlation coefficient. This doesn't mean you should necessarily exclude them, but you should be suspicious of them!

1. Sketch two scatterplots. In the first, include an outlier that would *inflate* the correlation between your two variables. In the second, include an outlier that would *deflate* the correlation between the two variables.



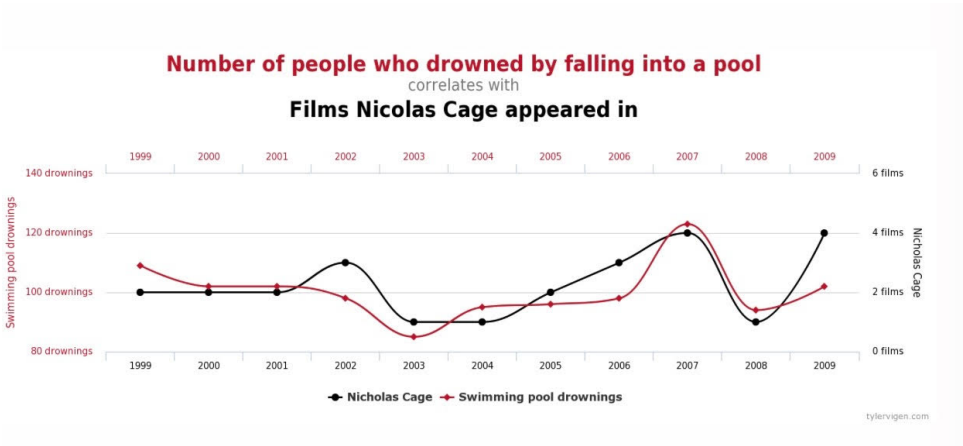


Confounding Variables:
We always need to be on the lookout for situations where there could be a third variable lurking in the background that is actually responsible for the variation in the response variable.

2. There is a very strong positive correlation between sunglass purchases and ice cream sales. Do you think that sunglass purchases are directly responsible for an increase in ice cream sales?

3. Can you identify a possible confounding variable in the above scenario?

Correlation vs. Causation:
Related to the last issue, we stress that just because there is a strong correlation between variables does NOT mean that one directly causes the other! We've all seen this in the news; data is used to make wild claims about variables that are actually unrelated. By drawing incorrect conclusions from our data, we could conclude that "ice cream sales increase murders", or, as seen below, we should expect more drownings whenever Nicholas Cage accepts a new movie role!



The takeaway here is: be careful not to jump to conclusions based solely on correlation. Look for confounding variables, or accept that it could be a simple coincidence. To prove causation, we need a carefully designed experiment with controls (coming up in Chapter 6!)

Extrapolation:
The final issue we need to be aware of is the dangers of extrapolating beyond the data. Let's investigate using our skin cancer mortality scatterplot.

4. Suppose we're interested in skin cancer mortality in Alaska (not currently on the plot). At the northernmost tip of Alaska, the latitude is 71.4. Predict the number of skin cancer deaths in this location.

5. Does your prediction make sense? What does this suggest about using regression lines for prediction?

Congrats! You are now regression experts. We'll investigate regression further in next week's lab and learn how to make R do computational part for us.