Lecture 20

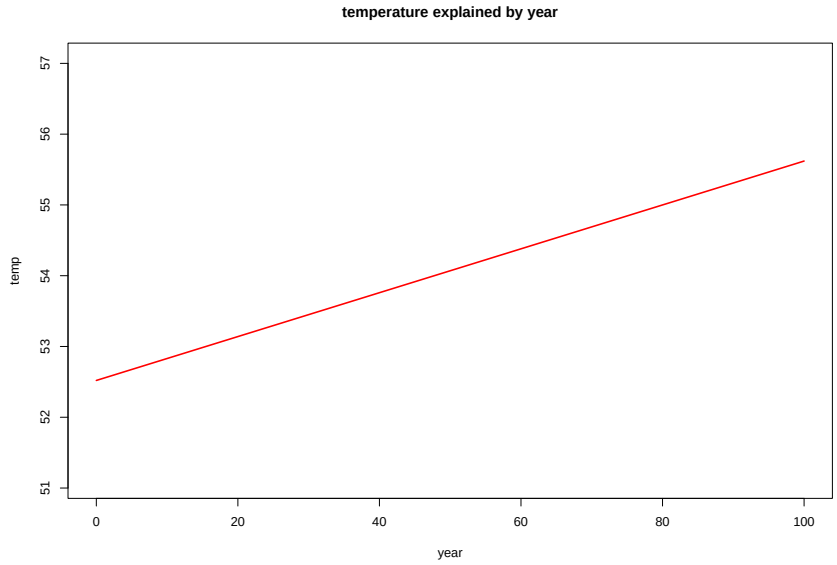
Cautions in Analyzing Associations

Extrapolation refers to using a regression line to predict y values for x values outside the observed range of data. This is riskier as we move farther from that range.

Example:
Extrapolation is seen most often in regression lines where the response variable is time and we want to make predictions into the future...values that are definitely outside our range of observed data.
Let's look at a regression line for average temperature (y) in central park predicted by time in years (x):

$$Temp = 52.52 + 0.031Year \tag{1}$$

Note that temperature is measured in Fahrenheit and year ranges from 0 to 100 where 0→1900 and 100→2000.



If the present trend continues what would we expect for the average temperature in central park for 2005:

Now, let's look much further into the future at the average temperature in central park in the year 3000:

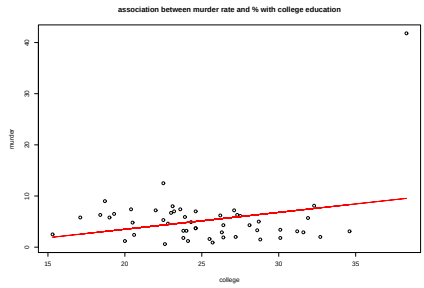
It's not sensible to assume that the same straight line trend will continue for the next 1,000 years based on information only from the past 100 years.

Remember to plot the data before you fit a regression line. This is not only to check for linearity (since we are only fitting linear regression lines) but also to check for **outliers**. One outlier can has a large effect on the results from regression analysis, an observation like this is called **influential.** For an observation to be influential, two conditions must hold:

- It's x value must be relatively high or low compared to the rest of the data
- It must fall quite far from the trend of the rest of the data

When both of these happen, the line tends to be pulled toward that data point and away from the trend of the rest of the points.

Example:
Let's look at 2003 data for the 50 states plus DC. We want if there is an association between murder rate (annual number of murders per 100,000 people) in a state and it's percentage of people with a college education.

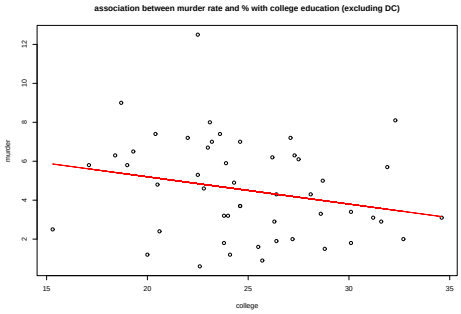


The regression equation for these variables for all 51 states is:

$$Murder = -3.1 + 0.33College \tag{2}$$

The slope is positive meaning that the predicted murder rate gets higher as college educated percentage increases. Does this mean that people get more violent through college education? Look at the outlier: It's the data from Washington, DC which has a murder rate of 41.8 and 38.3% of people have a college education. Is this an outlier?

Let's look at the data without the DC observation and see if there is a different trend.



$$Murder = 8.0 - 0.14College \tag{3}$$

CORRELATION DOES NOT IMPLY CAUSATION

In the last example we found that murder rate and education werepositively correlated when we looked at all 51 observations, but that does not mean that having a high level of education causes a high murder rate. Whenever we observe a correlation between variables x and y, there may be a third variable correlate with both x and y that is responsible for their association.

Example:
The "Gold Coast" of Australia, south of Brisbane, is famous for its magnificent beaches. Because of strong rip tides, however, each year many people drown. Data collected each month show a positive correlation between $y =$ number of people who drowned in that month and $x =$ number of gallons of ice cream sold in refreshment stands along the beach in that month.

Identify another variable that could responsible for this association.

Example:
Do tall students tend to have better vocabulary skills than short students? We might think so looking at a sample of students from grades 1, 6, and 12 of Lake Wobegon school district. The correlation was 0.81 between their height and their vocabulary test score: Taller students tended to have a higher vocabulary score.

Identify another variable that could responsible for this association.