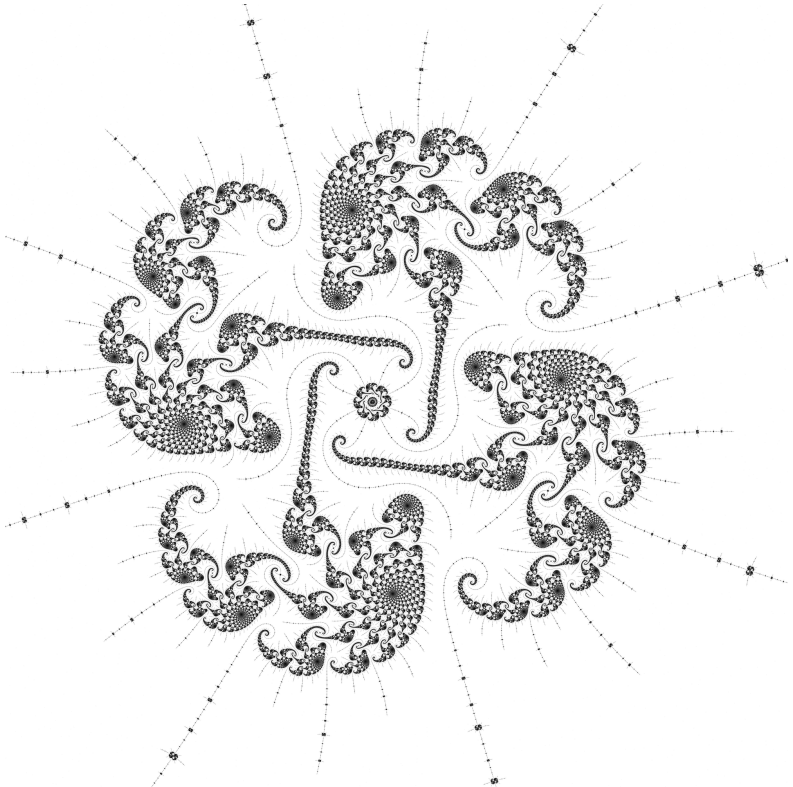
Correlation and Regression

Javier E. Flores

February 4, 2019

Mathematicians are Weird

Does anyone know what this is?



Mathematicians are Weird

This image is a visual representation of a Mandelbrot set, of which there are several variations.

The mathematical study of Mandelbrot sets began with work by mathematicians Adrien Douady and John H. Hubbard, but the name "Mandelbrot" is derived from another mathematician, Benoit Mandelbrot, for his highly influential work in fractal geometry.

The first visualization of a Mandelbrot set was drawn in 1978 by two other mathematicians, Robert W. Brooks and Peter Matelski.

Alex Grey

Looking at Mandelbrot sets, I can't help but think about artwork by Alex Grey, whose work was featured by the band *Tool* on several of their albums.

Math and LSD

Alex Grey has openly written about the influence of lysergic acid diethylamide (LSD) on his artistic vision.

This being said, I wonder whether Alex Grey and the mathematicians behind Mandelbrot sets had similar influences...

So to kick off today's lecture, I'll ask the question:

"Does LSD improve your mathematical ability?"

Math and LSD

As it so happens, a study published in 1968 by *Clinical Pharmacology and Therapeutics* was interested in this very question.

In the study, seven volunteers were intravenously administered doses of LSD.

Blood samples were then drawn over several time points and concentration levels of the drug were measured for each sample.

After each blood sample, the volunteers were given one of a series of equivalent tests in basic arithmetic.

Math and LSD

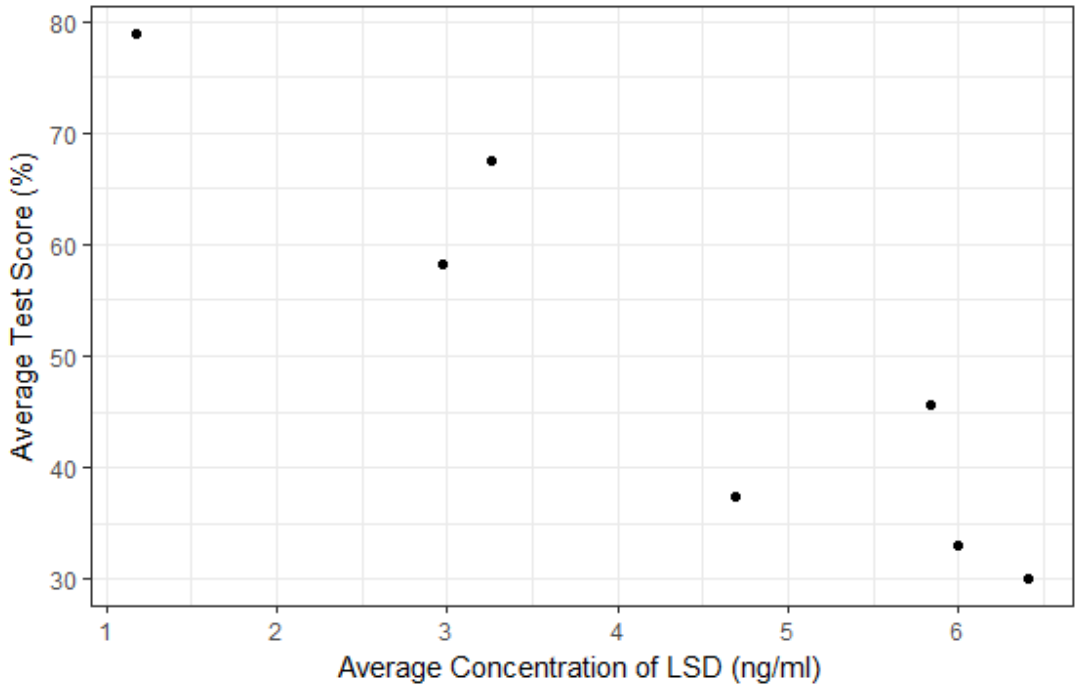
The table below displays the average LSD concentration and test results for each subject.

Subject	LSD concentration	Test Score
1	1.17	78.93
2	2.97	58.20
3	3.26	67.47
4	4.69	37.47
5	5.83	45.65
6	6.00	32.92
7	6.41	29.97

Scatterplots

A quick glance over the table might suggest an inverse relationship between LSD concentration and test performance.

Making determinations like this is not always possible (think large datasets), so we often use **scatterplots** to quickly visualize relationships between two quantitative variables.



Standardization

By these data, it's clear that LSD doesn't improve mathematical ability. In fact, the exact opposite seems to happen!

Given this observation, we might next be wondering about the strength of this association: "Just how bad is your math on LSD?"

Just eyeballing the trend, it would appear that the association is strong. However as statisticians, we would really like to somehow *quantify* the strength.

Before we can do this, it's important that we first learn about standardization.

Standardization

Standardization is a way for statisticians to transform variables so that they are "on an equal playing field". In other words, the variables are changed so that they can be directly compared.

To accomplish this, we compute z-scores for each variable.

$$z_i = \frac{x_i - \bar{x}}{s_x}$$

Here z_i is the **z-score** for the i^{th} case, $\bar{x}$ is the sample mean, and s_x is the sample standard deviation of the variable x (e.g. score or lsd concentration).

Why Standardize to Begin With?

Suppose you're told that the concentration of urea in your blood is 50 mg/dl above average. What do you conclude?

It's difficult to say:

What exactly is the average?

- If large, say 50,000 mg/dl, then there might be no cause for worry.
- If small, say 5 mg/dl, then you might be in trouble.

Why Standardize to Begin With?

Suppose now you're told that the concentration of urea in your blood is above the average by 4 times the sample standard deviation. What do you conclude?

In this case, the message is clearer - you need to see a doctor!

Remember that the standard deviation describes how spread out your data are. It can be thought of as an average distance from the mean.

To say that you are four standard deviations above the mean is to say that you are greater than the mean by four times the average distance in your sample!

In using standardization, non-experts are better able to interpret quantitative variables.

How Do Z-Scores Standardize?

Computing z-scores transforms variables so that they have a mean of 0 and standard deviation of 1.

As a result, the variable's original unit of measurement is removed.

Each Z-score may then be interpreted as the distance (in standard deviations) from the original variable's mean.

This might become apparent after studying the formula, but, as an example, let's compute the z-score for subject three's lsd concentration:

$$z_3 = \frac{x_3 - \bar{x}}{s_x} = \frac{3.26 - 4.33}{1.94} = -0.55$$

Practice

Download the Election Margin dataset and load it into Minitab. These data cover all US re-election races since 1940 and include the year, incumbent, incumbent's approval rating, incumbent's margin of victory or defeat, and the election result.

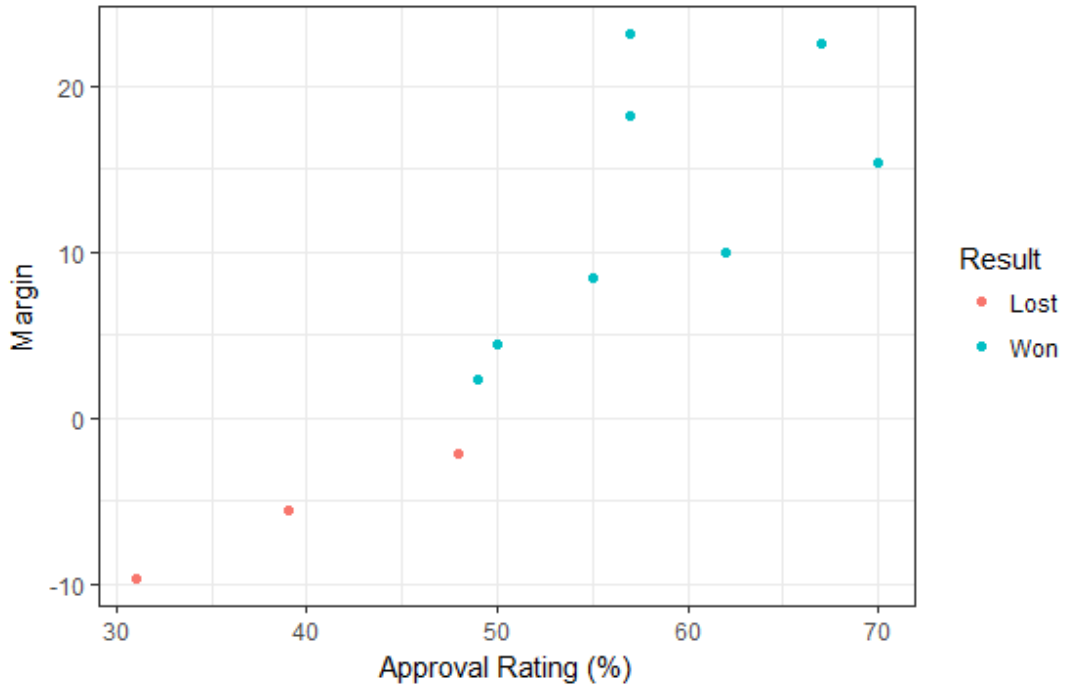
With your groups,

- Create a scatterplot between approval rating and margin of victory/defeat. Do you see a relationship? What approval rating appears necessary to win reelection?
- George W. Bush won reelection with a 49% approval rating. Calculate and interpret a z-score comparing Bush's approval rating to all other incumbents. Repeat this calculation using only those who won reelection.

Solution: Scatterplot

There appears to be a strong and positive relationship between approval rating and margin of victory.

Differentiating between those incumbents who won and lost, we see that ∼50% approval rating seems necessary for victory.



Solution: Z-scores

In the first case, we compute the z-score using the mean and standard deviation of *the entire sample*:

$$z_{bush} = \frac{x_{bush} - \bar{x}}{s_x} = \frac{49 - 52.9}{11} = -0.35$$

In the second case, we compute the z-score using the mean and standard deviation of *the incumbents who won reelection*:

$$z_{bush(re)} = \frac{x_{bush} - \bar{x}_{(re)}}{s_{x(re)}} = \frac{49 - 57.3}{7.6} = -1.1$$

Back to LSD

Now that we have an understanding of standardization and z-scores, we might next be wondering how they play a role in answering our original question:

> "Just how bad is your math on LSD?"

We can answer the above question using **Pearson's correlation**, otherwise known as the **correlation coefficient**:

$$r_{xy} = \frac{1}{n-1} \sum_{i}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

The formula for the correlation, r_{xy}, between two variables, x (e.g. LSD conc.) and y (e.g. score), is the average product of their z-scores!

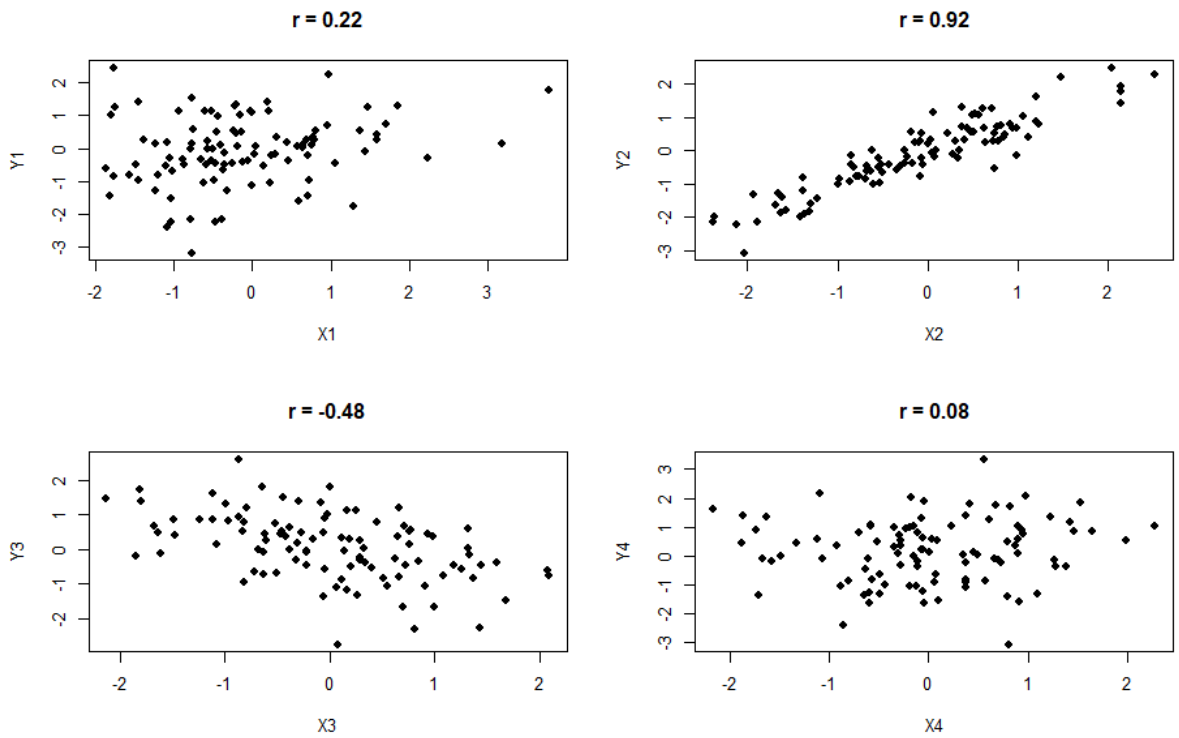
The Correlation Coefficient

Ranging between -1 and 1, the correlation coefficient is a summary statistic that quantifies the *strength* and *direction* of a linear association between two quantitative variables.

To denote the sample correlation, statisticians often use the lower case r. When talking about population correlation, the Greek letter 'rho' (ρ) is used.

The Correlation Coefficient

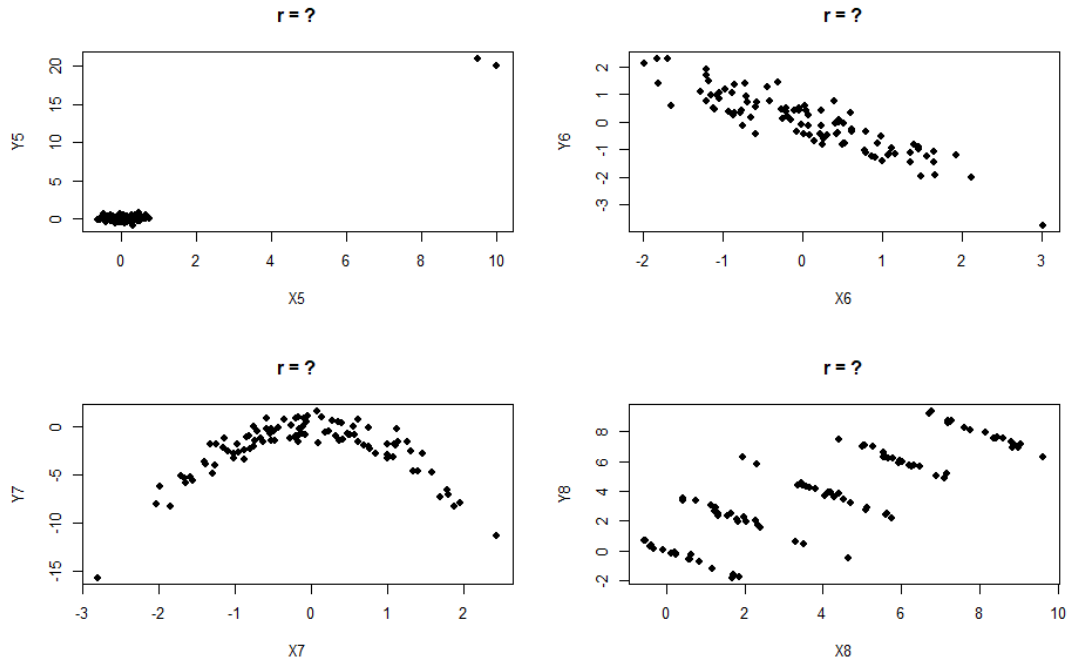
Shown below are a few scatterplots and their associated correlations.



Practice

With your groups,

- Compute Pearson's correlation for the LSD data. Is the correlation an appropriate measure of the strength of the association?
- Match the correlations 0.97, -0.9, 0.07, and 0.82 to the figures below. State whether using correlation to describe the relationship is appropriate.



Solution: Pearson's Correlation

The table below provides the z-scores for each variable in the LSD dataset.

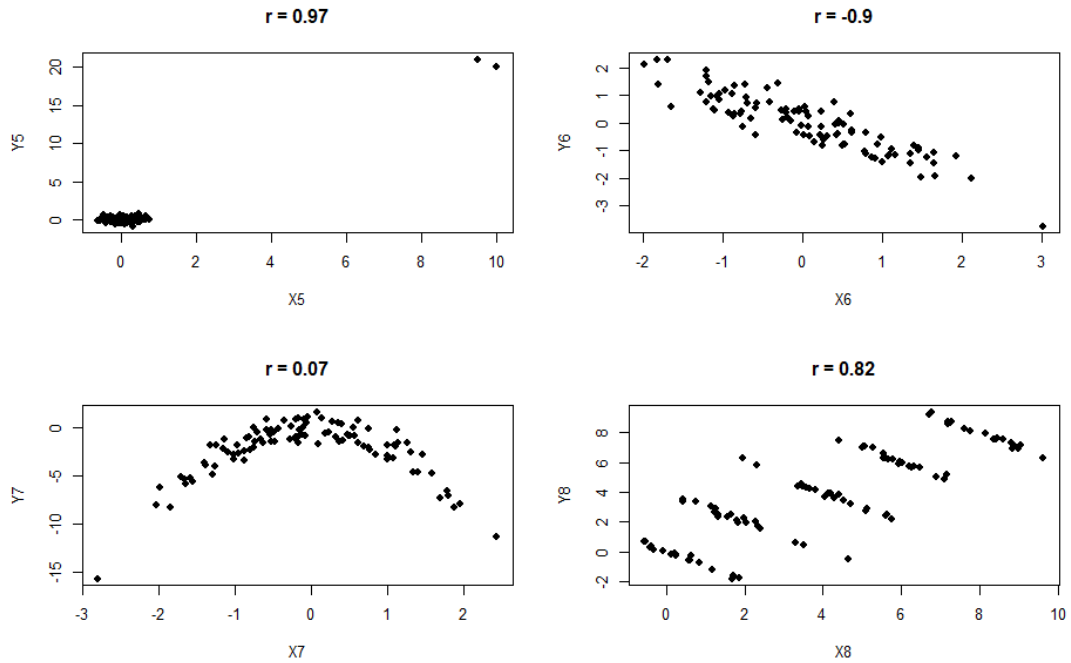
Subject	**LSD**	Z_{LSD}	**Test Score**	Z_{test}	$Z_{LSD} * Z_{test}$
1	1.17	-1.63	78.93	1.55	-2.53
2	2.97	-0.70	58.20	0.44	-0.31
3	3.26	-0.55	67.47	0.93	-0.52
4	4.69	0.18	37.47	-0.68	-0.13
5	5.83	0.77	45.65	-0.24	-0.18
6	6.00	0.86	32.92	-0.92	-0.79
7	6.41	1.07	29.97	-1.08	-1.16

Adding the product, $Z_{LSD} * Z_{test}$, across all rows and dividing by 6, we obtain a correlation coefficient of -0.94. Since the association appears linear, using Pearson's correlation is appropriate.

Solution: Correlation Matching

Only in the top right figure is Pearson's correlation an appropriate descriptor of the association.

There are clear associations in the bottom figures. However, the correlation coefficient is appropriate only for linear associations.



Ecological Correlations

The previous exercise provided some examples where computing a correlation could lead you astray.

Each of the previous examples emphasized the need for linear associations when computing and interpreting Pearson's correlation.

Another important distinction among correlations is whether they are **ecological**, or computed in the aggregate.

Neglecting this aspect of correlation may also lead to erroneous conclusions. As an example, let's consider a 1950 study demonstrating this concept.

Ecological Fallacy

In the study, the relationship between nativity and literacy in the United States was investigated.

The percent of the population that were foreign-born and the percentage who were literate were computed for each of the 48 states in the 1930's USA.

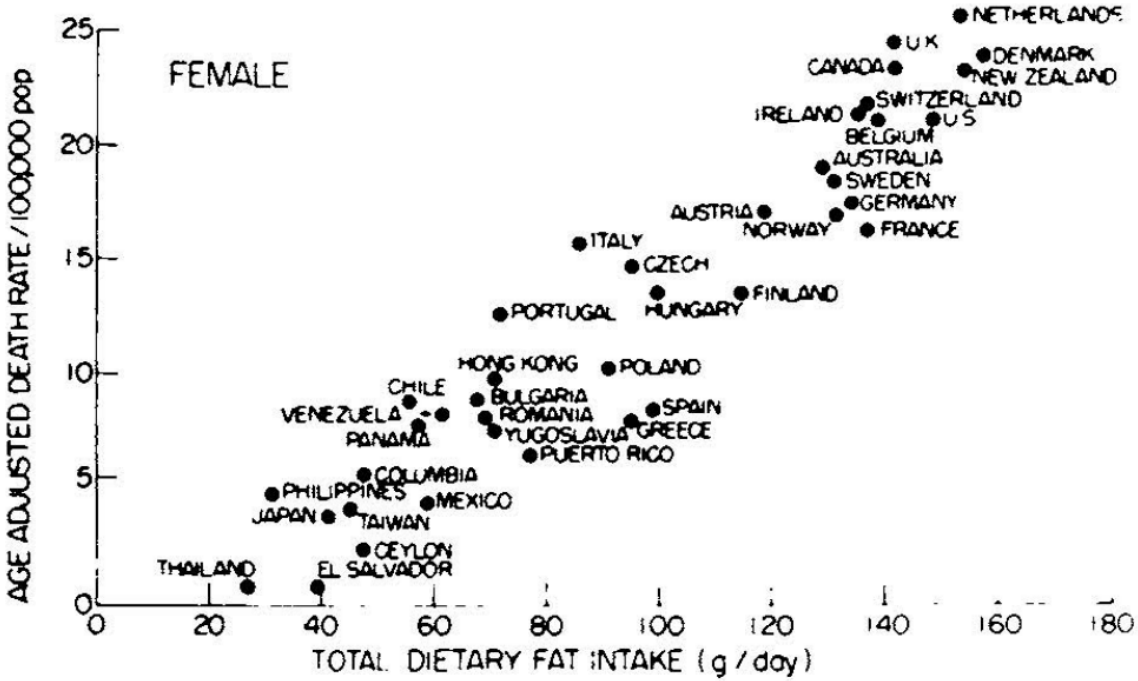
The correlation between these two variables was found to be 0.53, suggesting that foreign-born individuals were more likely to be literate.

However, the reality was exactly the opposite. If the correlation was computed at the individual level, as opposed to the ecological (state) level, the correlation was -0.11.

Ecological Correlation

From an article by Carroll in *Cancer Research* (1975):

From an article by Carroll in *Cancer Research* (1975):



Ecological can be used to draw conclusions at the level of aggregation. Problems are introduced only when you ignore the aggregation and make conclusions about the individual.

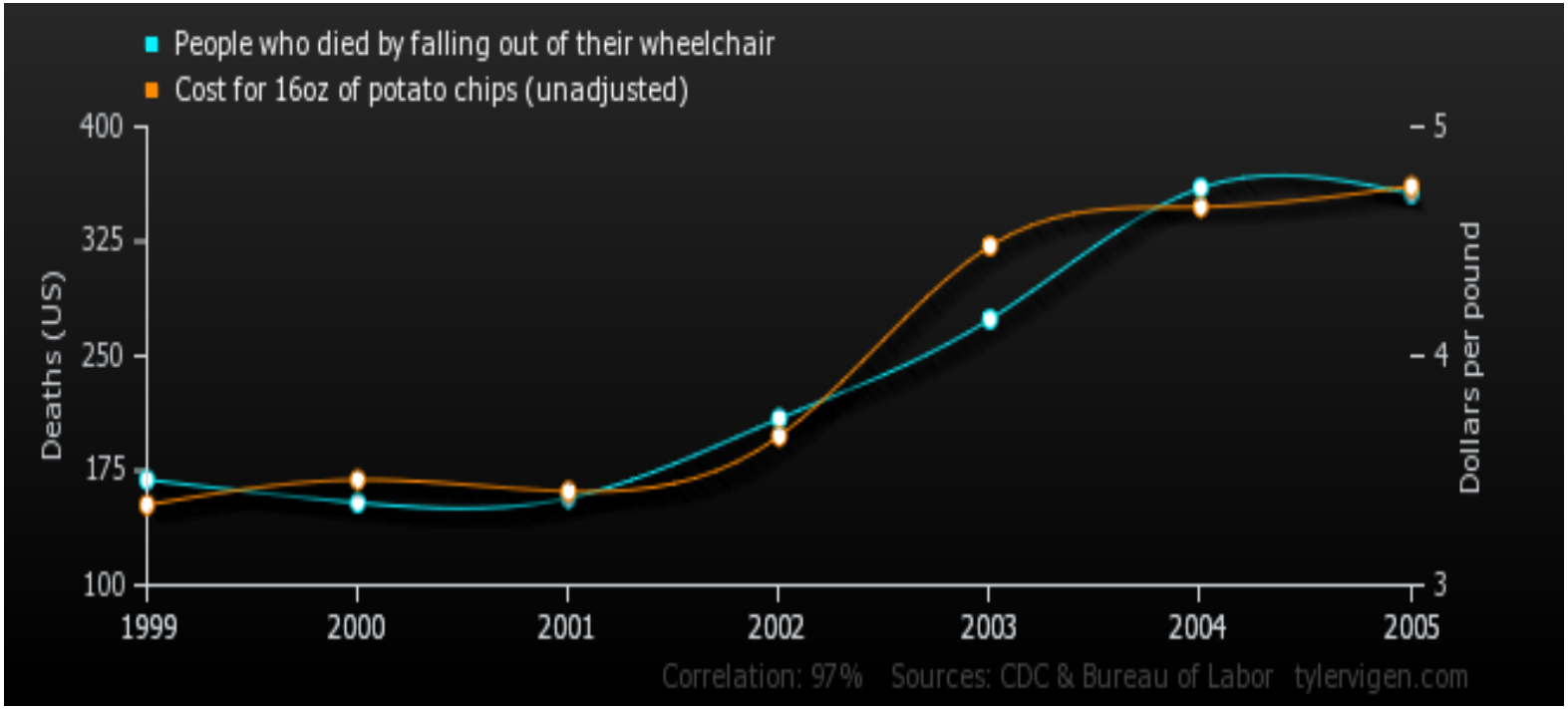
Correlation = Causation?

Thus far we've learned that so long as an association is linear and the appropriate distinction between aggregate and individual level data is made, we can use correlation to characterize the strength of an association.

Knowing that some associations are causal, can the correlation be used to support causal claims?

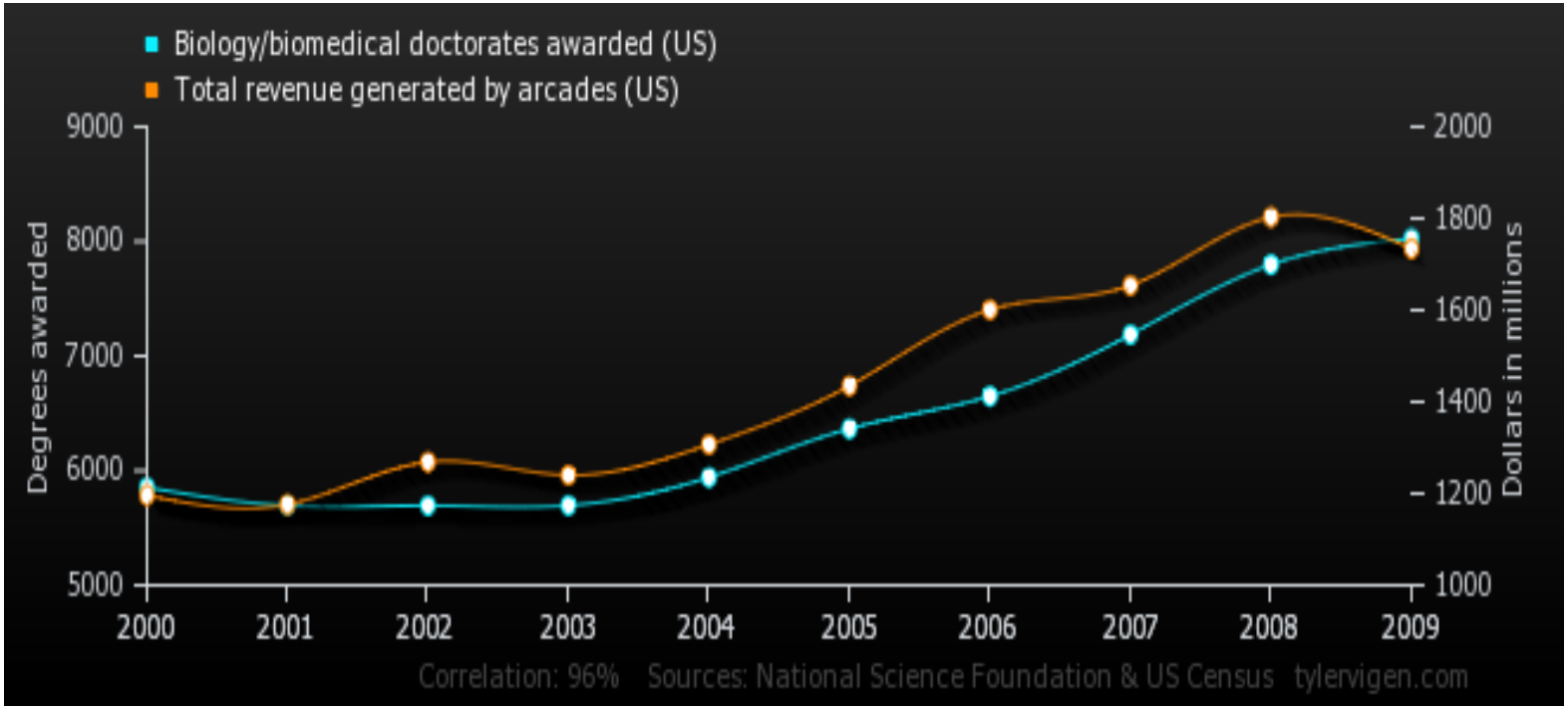
Consider the following three correlations...

Correlation = Causation?



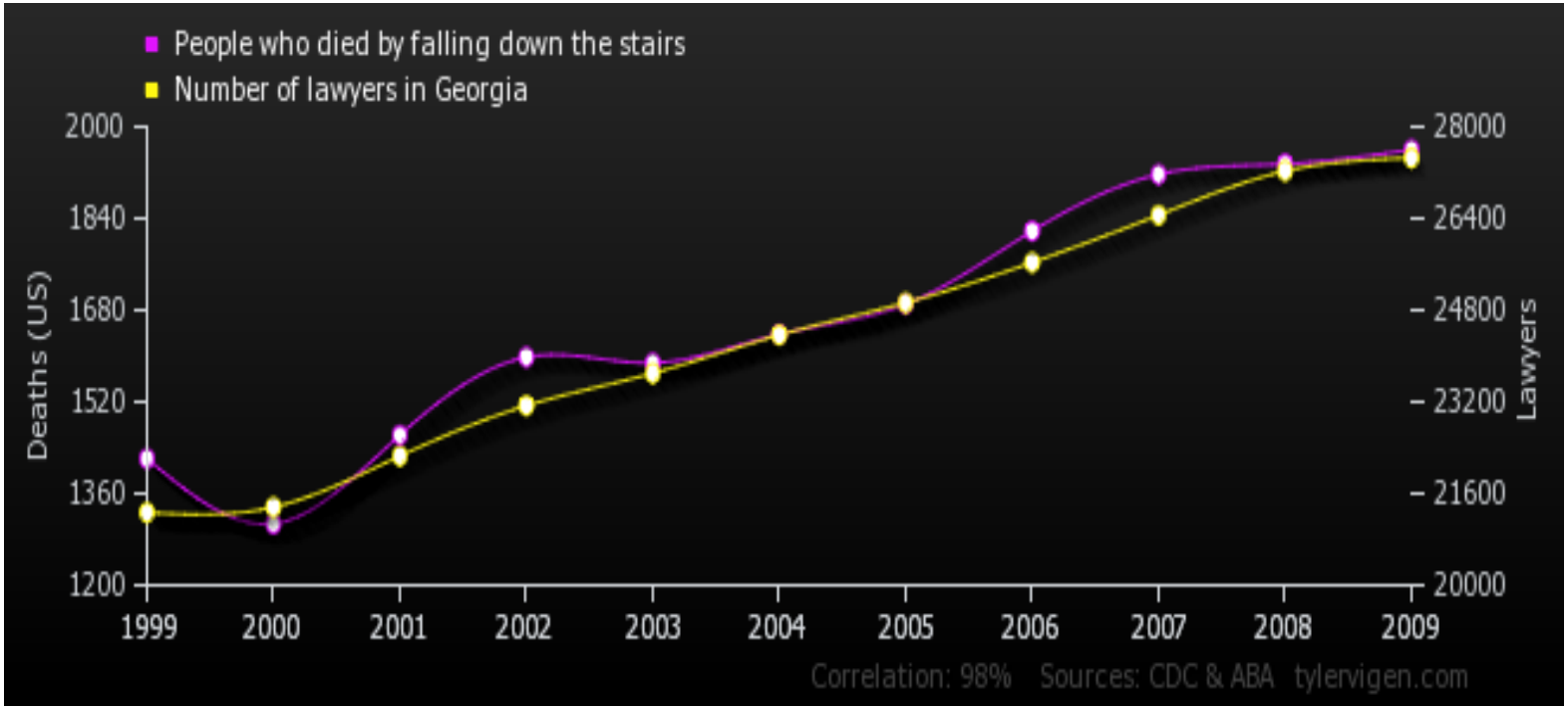
Will reducing the cost of potato chips save lives? ($r = 0.97$)

Correlation = Causation?



Do arcades empower biology/biomedical students to earn their PhD? ($r = 0.96$)

Correlation = Causation?



Are Georgian lawyers part of a secret cabal of assassins who kill only with stairs? ($r = 0.98$)

Correlation $\neq$ Causation

Obviously each of these questions are completely ludicrous, and each correlation even more so.

The point of sharing these *spurious* correlations is simple:

Correlation does NOT imply causation.

Feel free to browse the website that I pulled these examples from. You can find even wackier correlations if you look hard enough.

Correlation for Prediction

Even without granting the ability to talk about causality, correlations are still useful. One way that correlations may be used is for *prediction*.

Thinking back to our LSD dataset, we've already learned that LSD and math don't mix well, and the association between the two is strong.

This being said, what if we were interested in predicting different test scores based on different concentrations of LSD?

Correlation for Prediction

Suppose that for a certain subject the LSD concentration is 2.40 ng/dl, which is equivalent to one standard deviation below the sample average. What score would you then predict?

Given the negative relationship between LSD concentration and score, it would be reasonable to predict a score higher than the sample average. But by how much?

Since the LSD concentration is one standard deviation below average, should we predict a score one standard deviation above average?

Correlation for Prediction

Remember that score and concentration are not perfectly correlated. Rather, the computed correlation is -0.94.

If these two variables were perfectly correlated, then it would be entirely appropriate to predict a score one standard deviation above the sample average.

Otherwise, we need to adjust our prediction by the computed correlation.

Since our sample standard deviation for score is 18.61, we should predict a score of (-0.94)*(-18.61) + 50.09 = 67.58.

Correlation for Prediction

Generalizing this example, we can make predictions for correlated variables X and Y by following these steps:

1) Compute the z-score for the explanatory variable, X. In the previous example, $z_x = -1$.
2) Multiply the z-score by the correlation computed between X and Y in order to obtain the predicted z-score for the response variable, Y (i.e. $z_y = z_x * r_{xy} = -1 * -0.94$).
3) Unstandardize the predicted z-score by multiplying by the response standard deviation and adding the response mean (i.e. $y_{pred} = \bar{y} + z_y * s_y$).

Practice

In the late 19^{th} century, sir Francis Galton (the father of correlation), collected data on the heights of fathers (X) and their fully grown sons (Y) to determine whether they were associated.

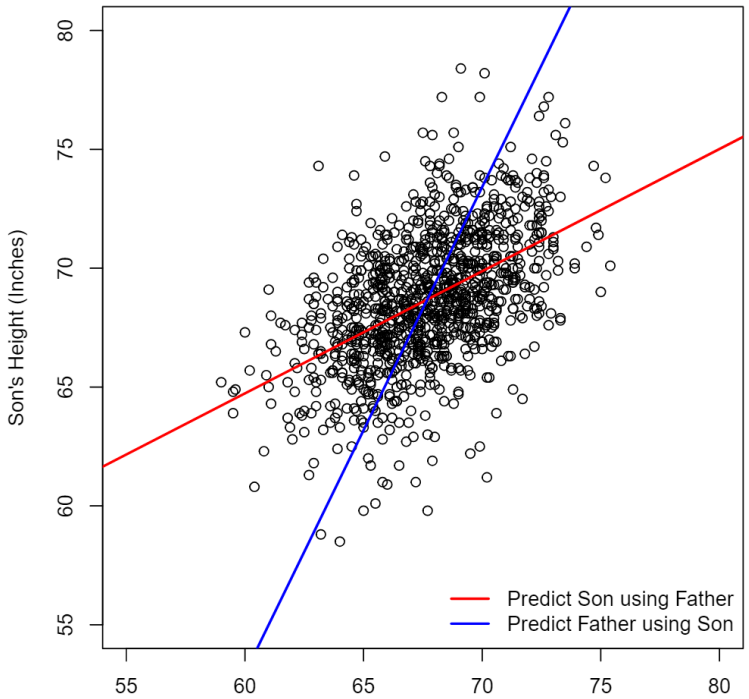
For his analysis, Galton computed the following statistics:

- $\bar{y} = 68.7$
- $\bar{x} = 67.7$
- $s_y = 2.8$
- $s_x = 2.7$
- $r_{xy} = 0.5$

With your group, predict the *son's* height of a father who is 65 inches tall. Next, predict the *father's* height of a son who is 67.3 inches tall.

Asymmetric Predictions!

We predicted the son of a 65-inch tall father to be 67.3 inches. However the predicted height of the father of a son who is 67.3 inches was 67 inches, and not 65 inches.



Symmetry and Asymmetry

The previous figure demonstrates that when making predictions, the choice of explanatory and response variables matters.

For this reason, we say that using correlation to make predictions is an **asymmetric** statistical method.

On the other hand, simply computing the correlation is a **symmetric** statistical method since $r_{xy} = r_{yx}$.

Regression

Regression is another asymmetric statistical method that may be used for prediction.

However, unlike with correlation-based predictions, regression allows us to quickly and easily determine how much our response variable Y may change with a change in our explanatory variable X.

In general, regression lines have the form:

$$\hat{y} = a + b * x$$

Regression allows us to **model** our response as a linear function of X. The **slope coefficient** (b) in this equation quantifies the predicted change in Y associated with changes in X.

Regression

Using the data Galton collected, we can fit a regression line to predict son's heights (Y) from the height of their fathers (X):

$$\hat{y} = 33.9 + 0.51 * x$$

Using this regression line, we can predict the height of any son by plugging in any given father's height.

We also see that $b = 0.51$, which means that we expect a 0.51 inch increase in son's height for every 1 inch increase in father's height.

Regression - Correlation Connection

If you're really keeping tabs, you may have noticed that the values for b and r_{xy} are extremely similar. Recall that $r_{xy} = 0.50$ and $b = 0.51$ for the Galton data.

The similarity between these two statistics is no coincidence. In fact, this is explained by a well-known result in statistics:

$$b = r_{xy} \frac{s_y}{s_x}$$

Since, in our Galton data, $s_y = 2.8$ and $s_x = 2.7$ are very close in value, the same bears out between r_{xy} and b.

Practice

Using the LSD dataset,

Subject	LSD concentration	Test Score
1	1.17	78.93
2	2.97	58.20
3	3.26	67.47
4	4.69	37.47
5	5.83	45.65
6	6.00	32.92
7	6.41	29.97

1) Use the correlation coefficient computed earlier to compute the predicted test score for subject six. How close is the prediction to the actual value?

2) Use Minitab to fit a regression line to these data. Use the regression equation to make the same prediction as above. Compare the predictions. Which is closer?

Solution

After standardizing our lsd concentration, we obtain a z-score of 0.86. Then, we can use the correlation, response variable sample mean, and response standard deviation to obtain our prediction:

$$\begin{aligned} y_{pred} &= \bar{y} + z_y * s_y \\ &= \bar{y} + z_x * r_{xy} * s_y \\ &= 50.09 + 0.86 * (-0.94) * (18.61) = 35.05 \end{aligned}$$

Taking the difference between this predicted value and the actual value, we obtain $35.05 - 32.92 = 2.13$.

Solution

We obtain the following regression equation through Minitab:

Regression Equation

Score = 89.12 - 9.01 LSD

We plug in the explanatory variable value for "LSD" in the equation above to get our prediction:
$89.12 - 9.01 * 6 = 35.06$

In comparing the predictions, we see that they are practically equal.

The Drake Curse

Despite his overwhelmingly successful rap career, Drake Graham seems to be haunted by a curse that stirs hate (and sometimes joy) in the hearts of sports fans everywhere.

Almost every single athlete or franchise that Drake decides he's a fan of meets some terrible fate.

As evidence of this curse, I provide you with three cases:

Case 1: Serena Williams

Arguably one of the greatest athletes in tennis, Serena Williams was named *Sports Illustrated*'s Sportsperson of the Year in 2015.

That same year, Serena was the favorite to be the first female tennis player to win a Grand Slam since 1998.

Then came Drake.

A few weeks after being seen kissing Drake in August 2015, Serena lost her bid for the Grand Slam after a defeat by Roberta Vinci (who was ranked 43^{rd} in the world at that time).

And to no one's surprise, Drake happened to be attending the game.

Case 2: Golden State Warriors

Over the past few years, the Golden State Warriors have become an NBA powerhouse.

Despite their incredible success, they, on at least one occasion, have been victims of the Drake curse.

On December 17, 2016 - only four days after GSW lost to the Milwaukee Bucks ending a 28 game winning streak - Drake was spotted third-wheeling Steph Curry (a star player for GSW) and his wife Ayesha at an In-N-Out.

Since you have to be *really* good friends with a couple to be comfortable enough to third wheel, Steph, Ayesha, and Drake must have been friends prior to the In-N-Out sighting.

Case 3: Johnny Football

Johnny Manziel, otherwise known as Johnny Football, was a highly sought after NFL recruit during the 2014 NFL Draft.

Manziel broke numerous NCAA Division 1 FBS and SEC records, and was the first freshman to win the Heisman Trophy, Manning Award and the Davey O'Brien National Quarterback Award.

But alas, the Drake curse struck once more.

Johnny, who was total bros with Drake by the time of draft, experienced the sad fate of being drafted by the Browns (one of the worst NFL teams).

Following the draft, his NFL debut was horrible and Manziel has since fallen (far) from the pedestal he had once stood.

Bonus: Super Bowl LIII

The Drake Curse

Do you believe in the Drake curse? Why?

Regression Fallacy

The "Drake Curse" is not truly the result of a cursed rapper, but rather a perfect example of what statisticians call the **regression fallacy**.

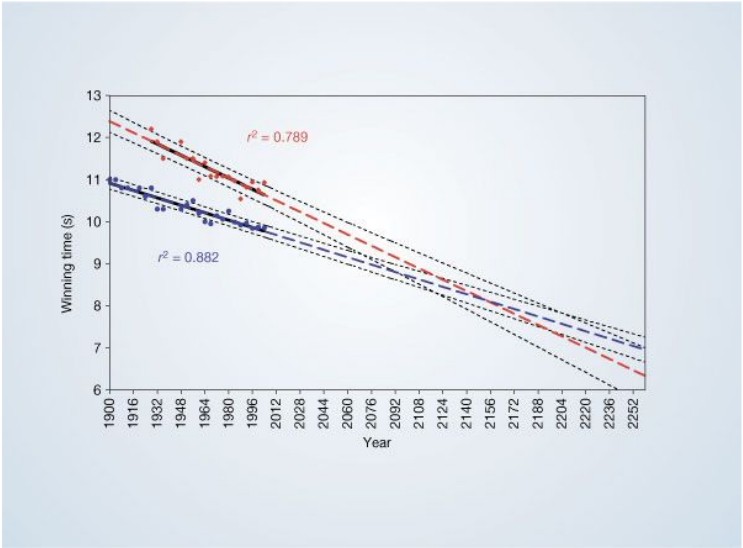
In each of these cases, Drake was what you call a "bandwagon" fan: he declared to be a fan only for those athletes/franchises that were doing *really* well.

While performance of each athlete is correlated from game to game, this correlation is certainly not 1.

As a result, these teams/athletes were bound to experience loss and *regress* to their performance average - regardless of whether Drake claimed to be a fan.

2156 Olympics

The figure below as taken from an article published in *Nature* titled "Momentous sprint at the 2156 Olympics?"

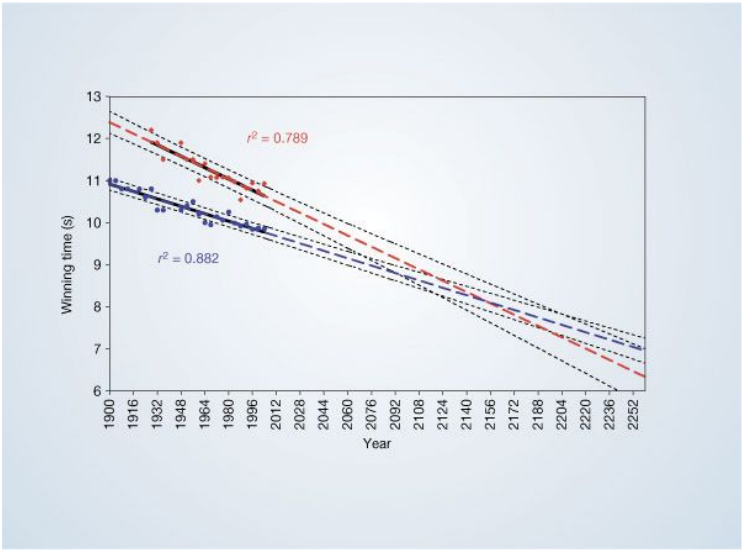


The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.

2156 Olympics

Based on the regression lines fitted to the winning times of the men's and women's 100m dash in every Olympics, the authors surmised that, given the observed trends, the 2156 Olympics would have women outsprint men.

Is this a reasonable conclusion to make?



The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.

Extrapolation

Well if this sick burn (below) is any indication, then no.

"*Sir - A. J. Tatem and colleagues calculate that women may out-sprint men by the middle of the twenty-second century. They omit to mention, however, that (according to their analysis) a far more interesting race should occur in about 2636, when times of less than zero seconds will be recorded.*

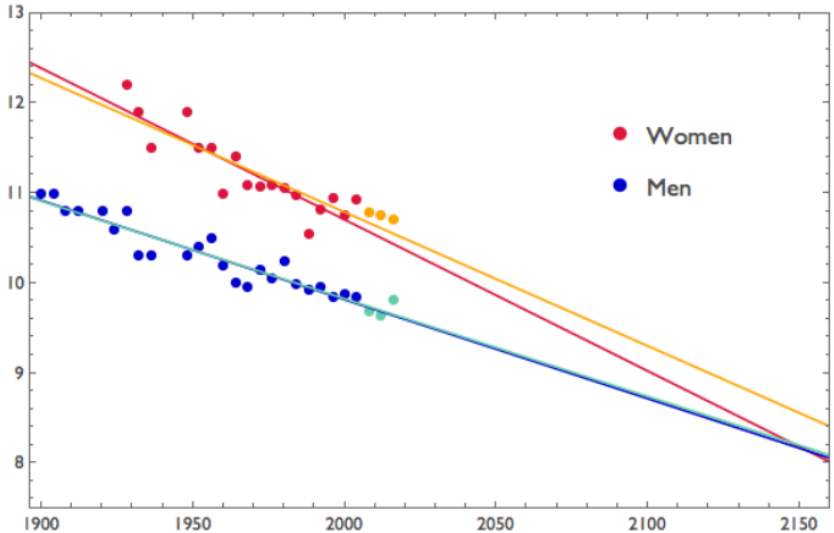
In the intervening 600 years, the authors may wish to address the obvious challenges raised for both time-keeping and the teaching of basic statistics."

Extrapolation

The lesson? Unless you want to get roasted by some savvy statistician, don't **extrapolate**, or predict beyond the observed range of your explanatory variable.

Since the original *Nature* article in 2004, there have been a few additional Olympics. Looking at the first three, we obtain the figure below.

Since the *Nature* paper was published, we've had three additional Olympic games. It is interesting to add the results from those three games (yellow and green points below) and see how the model has performed.



Regression

Now that we have some initial understanding of regression and its limitations, it might be worthwhile to discuss how exactly one obtains a regression line.

Previously we mentioned the relationship between the regression slope and the correlation coefficient.

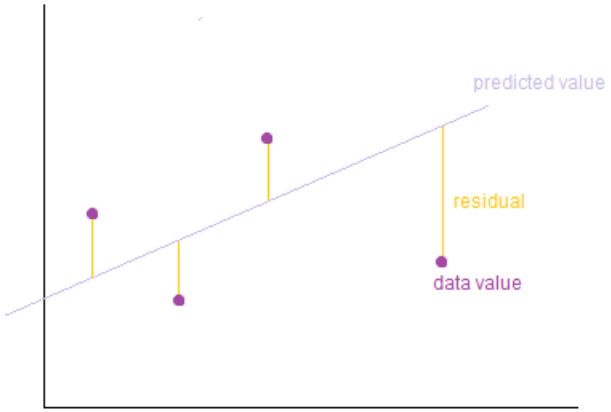
What we've yet to discover is how the entire line is obtained (i.e. the slope *and* intercept).

Regression

Regression lines are obtained by minimizing the **residual error** between the fitted regression line and your observed data points.

The residual error, or **residual**, for a given data point y_i is simply the difference between the data point and the predicted value $\hat{y}_i$ obtained from the proposed regression line.

On a scatterplot, the residuals are the vertical deviations from the observed data to the regression line.



Least Squares

Minimizing the residual error is equivalent to minimizing the following function:

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

where $\hat{y}_i = a + bx_i$. Finding the a and b which minimize this function is how the regression line is obtained.

Wrap-Up

Right now, you should...

- Be able to calculate and interpret z-scores.
- Be able to calculate, interpret, visualize, and know the limitations of correlation coefficients.
- Understand how correlation may be used for prediction.
- Be able to interpret and know the limitations of regression.
- Understand how regression relates to correlation and how it is used for prediction.

These notes cover sections 2.5 - 2.6 in the textbook. Please read through these sections and their examples along with any links provided in this lecture.