Not only is sensitive to outliers (it can go from negative to positive), but the very notion of association (e.g., when viewed through your eyes) is sensitive to outliers. In this example, even forgetting altogether, because of a single observation, the association goes from a negative one to a positive one. As such, the question of whether gun ownership and mass shootings are related does not have a yes/no answer at all! And it doesn't matter what measure of association you use.

There are some situations where even the scatterplot can fail in capturing all the facets of the relationship. E.g.

**Important:**
Association/correlation does not imply Causation.
Even if there is a strong correlation or association between 2 variables, that does not mean one causes the other. E.g. Shoe size and reading ability are associated. But I cannot increase my reading ability by wearing a larger shoe.

**Even more important:**
Even a non-causal association can be useful; for example, it can be used to predict one from the other. You can predict reading ability from shoe size.
Switching gears (even though it may not seem so, make sure you see! The change we are making: We are going from correlation to regression.)

What is an association between 2 vars. good for?

A

1) It can help in building theories.

2) It sets the stage for building predictive models, where one predicts one variable from the other.

Note: prediction is not in time.

Q

Can we use r itself for making predictions?

A

No. We need a fit, e.g. a line (i.e. regression model).

But you do not need a line for computing r.

Keep 2 things apart: Whereas r has (almost) nothing to do with the slope of anything, the fit does. After all, a fit is an equation e.g. \( y = \alpha + \beta x \), and so, the slope is very present.

BTW, in statistics, (line or curve) fitting is called **regression**

\[ y = \alpha + \beta x \]

Meaning: how much does \( y \) change with 1 unit change in \( x \)?

For finite points on a scatterplot, there are lots of possible fits. Which one do we pick?

\[ \text{Blood flow measured with Ultrasound} \]

\[ \text{HARD to measure} \]

\[ \text{Intracranial pressure (ICP)} \]
Called Ordinary Least Squares (OLS)

One very common selection criterion is to take the fit (line) that has the smallest \textbf{Sum of Squared Errors} (SSE)
or equivalently mean "", "", "", \( (MSE = \frac{1}{n} SSE) \)

Suppose we have \( n \) cases of data: \((x_i, y_i)\) \( i = 1, 2, 3 \ldots, n \)

\[ y = \alpha + \beta x \]

\[ y = \hat{y} \]

\[ \text{error} \]

\[ (x_3, \alpha + \beta x_3) \]

\[ (y_3, \hat{y}_3) \]

\[ \begin{align*}
MSE &= \frac{1}{n} SSE \\
&= \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2
\end{align*} \]

Minimize \( MSE \) \( \iff \) differentiate w.r.t. \( \alpha, \beta \) \( \iff \) set to zero \( \iff \) solve for the critical values of \( \alpha, \beta \) \( \Rightarrow \) \( \hat{\alpha}, \hat{\beta} \)

The specific values of \( \alpha, \beta \) that minimize \( SSE \) are called OLS estimates of \( \alpha, \beta \), and denoted \( \hat{\alpha}, \hat{\beta} \):

\[ \begin{align*}
\frac{\partial}{\partial \alpha} \text{MSE}(\alpha, \beta) \Big|_{\alpha = \hat{\alpha}, \beta = \hat{\beta}} &= 0 \\
\frac{\partial}{\partial \beta} \text{MSE}(\alpha, \beta) \Big|_{\alpha = \hat{\alpha}, \beta = \hat{\beta}} &= 0
\end{align*} \]
If you are not familiar with partial derivatives, \( \frac{\partial}{\partial \alpha} \), then just think of them as total derivatives. Let's do one:

\[
\frac{2}{\partial \beta} \text{MSE} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \beta} \left[ (y_i - \alpha - \beta x_i)^2 \right]
\]

\[
= \frac{1}{n} 2 \sum_{i=1}^{n} \left[ y_i - \alpha - \beta x_i \right] \frac{\partial}{\partial \beta} \left[ -x_i \right]
\]

\[
= -\frac{2}{n} \sum_{i=1}^{n} \left[ x_i y_i - \alpha x_i - \beta x_i^2 \right]
\]

\[
= -2 \left[ \frac{\sum_{i=1}^{n} x_i y_i}{n} - \alpha \frac{\sum_{i=1}^{n} x_i}{n} - \beta \frac{\sum_{i=1}^{n} x_i^2}{n} \right]
\]

\[
= -2 \left[ \bar{xy} - \alpha \bar{x} - \beta \bar{x}^2 \right]
\]

\[
\therefore \ (\bar{xy} - \alpha \bar{x} - \beta \bar{x}^2) = 0
\]

That's 1 eqn for 2 unknowns (\( \hat{\alpha}, \hat{\beta} \)). But there is \( \frac{\partial}{\partial \alpha} \):

\[
\frac{\partial}{\partial \alpha} \text{MSE} \bigg|_{\hat{\alpha}, \hat{\beta}} = 0 \Rightarrow \bar{y} - \hat{\alpha} - \hat{\beta} \bar{x} = 0
\]

See how, below.

Now we have 2 eqns for 2 unknowns. Solve!

\[
\hat{\beta} = \frac{\bar{xy} - \bar{x} \bar{y}}{\bar{x}^2 - \bar{x}^2}, \quad \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}
\]

Normal equation of regression.

\[\text{R: lm(y~x)}\]

Notation:

→ What I call \( \hat{\alpha}, \hat{\beta} \) are denoted a and b in book.

\( \text{SSE is } \sum \text{SSResid} \)

(I have good reasons for using my notation!)

→ The book also introduces the notation:

\[
S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad \text{Numerators of Sample Var.} \quad S_x^2, S_y^2,
\]

\[
S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})
\]

\[
S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2
\]

In which case it's easy to show that \( \hat{\beta} = \frac{\bar{xy} - \bar{x} \bar{y}}{\bar{x}^2 - \bar{x}^2} = \frac{S_{xy}}{S_{xx}} \).
Example

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>xy</th>
<th>x²</th>
<th>y²</th>
</tr>
</thead>
<tbody>
<tr>
<td>72</td>
<td>220</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>180</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>68</td>
<td>118</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>190</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Weight vs Blood Flow**

**Easy**

Joe: 70, 180
65, 120
68, 118
70, 190

\[ \hat{\beta} = \frac{\sum xy - \overline{x} \sum y}{\sum x^2 - \overline{x}^2} = \frac{11224.8 - 69(161.6)}{4766.6 - 69(69)} = 13.28 \]

\[ \hat{\alpha} = \overline{y} - \hat{\beta} \overline{x} = 161.6 - 13.28(69) = -7.55 \]

\[ \hat{\gamma}(x) = \hat{\beta}x + \hat{\alpha} = 13.28x - 7.55 \]

\[ \hat{\gamma}(70) = 13.28(70) - 7.55 = 914.9 \text{ pounds} \]

We can now predict everyone's y from their x.

For the people in the data set, we can also find their error/residual.

\[ y = \hat{\gamma} + (y - \hat{\gamma}) \]

For people outside the data set (e.g., Jane) we can predict their y from their x, but we cannot compute error, because we don't know their true y. In class, we'll address this issue.

Finally, be **warned** if you extrapolate.

\[ x = 0 \implies y = -7.55 \text{ pounds}! \]
Values of modulus of elasticity (MoE, the ratio of stress, i.e., force per unit area, to strain, i.e., deformation per unit length, in GPa) and flexural strength (a measure of the ability to resist failure in bending in MPa) were determined for a sample of concrete beams of a certain type, resulting in the following data (read from a graph in the article "Effects of Aggregates and Microfilers on the Flexural Properties of Concrete," Magazine of Concrete Research, 1997 8198):

MoE:
29.8 33.2 33.7 35.3 35.5 36.1 36.2 36.3 37.5 37.7 38.7 38.8 39.6 41.0 42.8 43.5 45.6 46.0 46.9 48.0 49.3 51.7 62.6 69.8 79.5 80.0

Strength:
5.9 7.2 7.3 6.3 8.1 6.8 7.0 6.6 6.8 5.7 6.3 7.9 9.0 8.2 8.7 7.8 9.7 7.4 7.7 7.8 7.7 11.6 11.3 11.8 10.7

a) Plot a scatterplot of Strength vs. MoE. By R.
b) Make a boxplot of MoE, and of Strength. By R.
c) Make a qqplot of MoE, and of Strength. By R.
d) Compute the correlation coefficient between MoE and Strength. By hand. You may use the computer to compute sample means of necessary quantities, but you must use one of the formulas for r.
e) Compare it with the correlation you get from cor() in R.
f) Compute the equation of the OLS fit (i.e., the intercept and slope). By hand. You may use the computer to compute sample means of necessary quantities, but you must use the formulas for OLS intercept and slope.
g) Interpret the slope.
h) Predict Strength when MoE is 39.0. By hand.
i) Compute the sum squared error (SSE, or SSR resid). By hand, but you may use R to compute sample means of necessary quantities.

I encourage you to come-up with other examples with "Easy" and "Hard" variables, because you're likely to come across something practically useful. But, even if you don't want to do that, do develop a regression model for predicting one of the continuous variables in your collected data from the other continuous variable. By R. Include your code, and report and interpret the slope parameter.