Interpreting Correlation and Regression

Case Study 19 (pg. 117 of reader)

Longitudinal data vs. cross-sectional data -
- Cross-sectional data is a snapshot of a sample of the population at one moment in time.
- Longitudinal data would be following people along at multiple time points.

Fact: It is very hard to draw valid longitudinal conclusions from cross-sectional data because PCF's like the year of birth in C.S. 19.

- With cross-sectional data, people in the snapshot differ from each other not only by x and y but also by PCFs (z).

(See pg. 121 of reader)

* Easier and cheaper to get cross-sectional data which is why people try to draw longitudinal data conclusions from cross-sectional data.

Association ↔ correlation ≠ causation because maybe

\[ x \rightarrow y \] or \[ y \rightarrow x \] or \[ x \rightarrow y \rightarrow z \] (associative causation)

Association is not causation, because of PCFs!

Ex. p.122
Regression in controlled experiments
- ex. Hooke's law in physics

\[ y = a + bx \]

"Stretch factor" - amount by which a 1kg weight would cause the spring to stretch

**Here you can infer causation since it is a controlled experiment**
when x is under experimental control and all other factors have been held constant, the slope in regression equation has a valid cause and effect interpretation.

Regression in observational studies
(\text{- cross-sectional (snapshot) data on n individuals (measure x, y on each person))}

Slope does not necessarily have a valid causal interpretation in this case, in fact often it doesn't have any causal meaning at all.

\begin{itemize}
\item \textbf{ex. 3} (Census is cross-sectional study)
\item \textbf{Variable} \hspace{0.5cm} \textbf{mean} \hspace{0.5cm} \textbf{SD}
\item \text{Ed. Level} \text{(x)} \hspace{0.5cm} 14 \text{ yrs.} \hspace{0.5cm} 3 \text{ yrs.} \hspace{0.5cm} n = 144
\item \text{Income} \text{(y)} \hspace{0.5cm} \$8,000 \hspace{0.5cm} \$3,000 \hspace{0.5cm} \text{(s_y)} \hspace{0.5cm} r = 0.4
\end{itemize}

(5 number summary - $\bar{x}$, $\bar{y}$, $s_x$, $s_y$, r)

Slope = $r \frac{S_y}{S_x} = (0.4) \frac{\$3000}{3 \text{ yrs.}} = \$400 \text{ income/yr. of education}$

Q: Does slope mean anything causally? No, because the causal conclusion on pg. 122 is an attempt to draw a longitudinal conclusion from cross-sectional data.
03. If that is not the meaning of the slope, what is it?
   "Associated with an increase of y is an increase of \( \text{slope} \) in x."

In other words, if I compare 2 groups of women whose education levels differ by 1 year, I expect to see their incomes differing by, on average, $400. This is a cross-sectional conclusion from cross-sectional data.