Lesson Plan

- Correlation
- Causation
- Regression
6.1 More About Correlation

The correlation coefficient, denoted by $r$, measures the strength of the linear association between $X$ and $Y$ values in a scatterplot. $r$ lies between -1 and 1, inclusive.

- $r$ equals 1 if all points lie on a line with positive slope.
- $r$ equals -1 if all points lie on a line with negative slope.
- $r$ lies between -1 and 1, inclusive.
- $r^2$ is the proportion of variation in $Y$ that is explained by $X$.

Non-zero $r$ does not imply a causal relationship.
Some examples of scatterplots: positive and negative correlation, outliers.
Correlations are often high when some factor affects both $X$ and $Y$. But sometimes, there might be a causal link. Hours of study is probably affected by lifestyle.

So it is hard to argue that correlation implies causation. GPA does not cause SAT, and Rob Zombie does not hurt GPA.

But sometimes, there might be a causal link. GPA does not cause SAT, and Rob Zombie does not hurt GPA. GPA is probably affected by IQ. Number of hours spent listening to Rob Zombie and GPA are both affected by lifestyle.

Correlations are often high when some factor affects both $X$ and $Y$. GPA and SAT scores are both affected by IQ.
Ecological correlations occur when X or Y or both is an average, proportion, or a percentage for a group. Here causation is especially difficult to show.

What kinds of confounding factors might the tobacco companies have suggested to explain Doll's finding?

The original link between smoking and lung cancer was an ecological correlation (Doll, 1955). The scatterplot showed the lung cancer rate against the proportion of smokers for 11 different countries. Note: A proportion is informationally equivalent to a percentage. And proportion, or a percentage, or both is an average, or a proportion when X or Y or both is an average,
A study showed that cantons in Switzerland that had high literacy rates also had high suicide rates.

What kind of studies would help determine whether a specific factor suggested by an ecological correlation is in fact causal?

Would this be an ecological correlation?

Suppose one plotted GPA against SAT score for a set of individuals. Is this an ecological correlation? What might be going on?

Would this be an ecological correlation?
Regression terminology:

The response variable is labelled $Y$. This is sometimes called the dependent variable.

The explanatory variable is labelled $X$. This is sometimes called the independent variable, or the covariate.

In multiple regression there are more than one kind of explanatory variable. For example, one might try to predict one’s grade in statistics using both high school GPA and IQ.
The book discusses the SD line. This is not the same as the regression line. This is not the same as the regression line. The SD line plots the values that are ±2, ±1, 0, ±1, ±2 standard deviations from the mean X. The SD line plots the values that are ±2, ±1, 0, ±1, ±2 standard deviations from the mean Y. This is one plotted height against weight, and your sample found that the mean height was 5’6” with an sd of 4” and the mean weight was 150 pounds with an sd of 20 pounds, then the SD line would contain (5’6”, 150), (5’10”, 170), (5’2”, 130), etc.
Regression tries to find the "best" straight line to the data. Specifically, it tries to fit the line that minimizes the sum of the squared deviations from each point to the line. Note: This does not measure deviation as the perpendicular distance from the point to the line.
Regression tries to predict the average value of $Y$ for a specific value of $X$. This is not the same as saying that an individual value lies on the line. In fact, an individual is often likely to be far from the line.

Beavis complains that he studied consistently for 10 hours each week, but his exam grade was only 70. Beavis says they should work for 10 hours a week on the material if they want to score a 90.

If you are advising a class on how to study, you tell them that the regression model says they should work for 10 hours a week on the material if they want to score a 90. Beavis is surprised that the regression line is $Y = 20 + 7X$. (Is this reasonable? When would it break down?)

For example, suppose we regress exam grade against number of hours of study per week. Assume the regression line is $Y = 20 + 7X$. This is not the same as saying that an individual value lies on the line.
Be aware that regressing weight as a function of height gives a different regression line than regressing height against weight. If your best estimate of the weight of a man who is 5'10" is 170 pounds, that does not mean that the best estimate of the height of a man who weighs 170 pounds is 5'10".

The book mentions the test-retest example, and also the father-son height example. The regression fallacy mistakenly argues that there is some effect or force that causes sons to be more average than their fathers. EFFECT VS. FALLACY. In fact, this is only the natural operation of random chance.

What can you say about the performance of baseball players in the first and second halves of the season? Or stock traders, or new employees?
The mathematical model for regression assumes that:

1. Each point \((X_i; Y_i)\) in the scatterplot satisfies:
   \[ Y_i = aX_i + b + \epsilon_i \]
   where the \(\epsilon_i\) have a normal distribution with mean zero and
   \(\epsilon_i \sim N(0, \sigma)\).

2. The \(\epsilon_i\) have nothing to do with one another. A large error does not tend to be followed by another large error, for example.

3. The errors \(\epsilon_i\) have nothing to do without error. Thus, all the error occurs in the vertical direction, and we do not need to minimize perpendicular distance to the line.
How does one find the estimates \( a \) and \( b \) of the coefficients in the regression equation? We need to get the values that minimize the sum of squared vertical distances.

The sum of the squared vertical distances is

\[
\sum_i (Y_i - aX_i - b)^2
\]

So, take the derivative of \( f(a, b) \) with respect to \( a \) and \( b \), set these equal to zero, and solve. One finds that:

\[
\begin{align*}
\hat{b} &= \bar{Y} - \bar{X}a \\
\hat{a} &= \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(x_i - \bar{x})^2}
\end{align*}
\]

The squared vertical distance is

\[
\sum_i (Y_i - (aX_i + b))^2 = (q', v)f
\]