Cautions about Correlation and Regression
A **residual** is the difference between an observed value of the dependent variable and the value predicted by the regression line.

\[
\text{residual} = \text{observed } y - \text{predicted } y
\]

\[
\text{residual} = y - \hat{y}
\]
Residuals

- Residuals show how far the data fall from our regression line.
  - They help us assess the fit of a regression line.
  - The mean of the least-squares residuals is always 0.

- A residual plot is a scatterplot of the regression residuals against the independent variable.
Outliers and Influential Observations

An **outlier** is an observation that lies outside the overall pattern of the other observations.

- Points that are outliers in the y direction of a scatterplot have large residual values.

An observation is **influential** for a statistical calculation if removing it would markedly change the result of the calculation.

- Points that are outliers in the x direction of a scatterplot are often influential for the least-squares regression line.
Beware!

- Correlation measures only linear association, and fitting a straight line makes sense only when the overall pattern of the relationship is linear.

- Extrapolation often produces unreliable predictions.

- Correlation and least-squares regression are affected by outliers and influential points.
Correlation based on averages

A correlation based on averages over many individuals is usually higher than the correlation between the same variables based on data for individuals.
Explaining association

Even when direct causation is present, it is rarely a complete explanation of an association between two variables.

Even well established causal relations may not generalize to other settings.
Warning!

Two variables are confounded when their effects on a response variable cannot be distinguished from each other.

Even a strong association between 2 variables is not by itself good evidence that there is a cause-and-effect link between the variables.

Review criteria on page 184