3.8: Putting It Together- Examining Relationships- Quantitative Data

Let's Summarize

- We use a *scatterplot* to graph the relationship between two quantitative variables. In a scatterplot, each dot represents an individual. We always plot the explanatory variable on the horizontal x-axis.
- When we explore a relationship between two quantitative variables using a scatterplot, we describe the overall pattern (*direction*, *form*, and *strength*) and deviations from the pattern (*outliers*).
- When the *form of a relationship is linear*, we use the correlation coefficient, *r*, to measure the strength and direction of the linear relationship. The correlation ranges between −1 an 1. If the pattern is linear, an *r*-value near −1 indicates a strong negative linear relationship and an *r*-value near +1 indicates a strong positive linear relationship. Following are some cautions about interpreting correlation:
  - **Always make a scatterplot before interpreting *r*.** Correlation is affected by outliers and should be used only when the pattern in the data is linear.
  - **Association does not imply causation.** Do not interpret a high correlation between explanatory and response variables as a cause-and-effect relationship.
  - **Beware of lurking variables** that may be explaining the relationship seen in the data.
- The line that best summarizes a linear relationship is the *least-squares regression line*. The least-squares line is the best fit for the data because it gives the best predictions with the least amount of error. The most common measurement of overall error is the sum of the squares of the errors, SSE. The least-squares line is the line with the smallest SSE.
- We use the least-squares regression line to predict the value of the response variable from a value of the explanatory variable. **Avoid making predictions outside the range of the data.** (This is called *extrapolation*.)
- We have two methods for *finding the equation of the least-squares regression line:* Predicted $y = a + b * x$
  - We use technology to find the equation of the least-squares regression line: Predicted $y = a + b * x$
  - We use summary statistics for x and y and the correlation. Using this method, we can calculate the slope b and the y-intercept a using the following:
- The *slope of the least-squares regression* line is the average change in the predicted values when the explanatory variable increases by 1 unit.
- When we use a regression line to make predictions, there is error in the prediction. We calculate this error as Observed value − Predicted value. This prediction error is also called a *residual*.
- We use *residual plots* to determine whether a linear model is a good summary of the relationship between the explanatory and response variables. In particular, we look for any unexpected patterns in the residuals that may suggest that the data is not linear in form.
- We have two numeric measures to help us judge how well the regression line models the data:
  - The square of the correlation, r^2 , is the proportion of the variation in the response variable that is explained by the least-squares regression line.
  - The standard error of the regression, s_e , gives a typical prediction error based on all of the data. It roughly measures the average distance of the data from the regression line. In this way, it is similar to the standard deviation, which roughly measures average distance from the mean.

Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 3.8: Putting It Together- Examining Relationships- Quantitative Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lumen Learning.