AP Statistics Chapter 3 – Examining Relationships

3.1: Scatterplots and Correlation

Explanatory and Response Variables
A **response variable** measures an outcome of a study. An **explanatory variable** attempts to explain the observed outcomes. The explanatory variable is sometimes referred to as the ***independent*** variable and is typically symbolized by the variable ***x***. The response variable is sometimes referred to as the ***dependent*** variable and is typically symbolized by the variable ***y***.

Scatterplot
A **scatterplot** shows the relationship between two quantitative variables measured on the same individuals. The values of the explanatory variable appear on the horizontal axis, and the values of the response variable appear on the vertical axis. If there is no clear explanatory/response relationship between the two variables, then either variable can be placed on either axis. Each individual in the data set appears as a single point in the plot fixed by the values of both variables for that individual.

Examining a Scatterplot
In any graph of data, look for patterns and deviations from the pattern. Describe the overall **pattern** of a scatterplot by the **form**, **direction** and **strength** of the relationship.

- **Form** can be described as **linear** or **curved**.
- **Direction** can be described as **positive** or **negative** or **neither**.
- **Strength** can be described as **weak**, **moderate** or **strong**.

A **deviation** from the overall pattern of a scatterplot is called an **outlier**.

Association

- Two variables are **positively associated** if as one increases the other increases.
- Two variables are **negatively associated** if as one increases the other decreases.

Correlation
Correlation measures the strength and direction of the relationship between two quantitative variables. Correlation is usually represented by the letter ***r***.

Facts about Correlation

1. When calculating correlation, it makes no difference which variable is x and which is y.
2. Correlation is only calculated for quantitative variables, not categorical.
3. The value of r does not change if the units of x and/or y are changed.
4. Positive r indicates a positive association between x and y. Negative r indicates a negative association.
5. Correlation is always a number between -1 and +1. Values close to +1 or -1 indicate that the points lie close to a line. The extreme values of +1 and -1 are only achieved when the points are perfectly linear.
6. Correlation measures the strength of a linear relationship between two variables, not curved relationships.
7. Correlation, like the mean and standard deviation, is nonresistant. Recall that this means that it is greatly affected by outliers.

3.2: Least-Squares Regression

Regression Line
A **regression line** is a straight line that describes how a response variable *y* changes as an explanatory variable *x* changes. The line is often to predict values of y for given values of x. Regression, unlike correlation, requires an explanatory/response relationship. In other words, when x and y are reversed, the regression line changes. Recall that correlation is the same no matter which variable is x and which is y.

Least-Squares Regression Line
The **least-squares regression line** is the line that makes the sum of the squares of the vertical distances from the data points to the line as small as possible.

Equation of the Least-Squares Regression Line
To find the equation of the regression line in the form $y = a + bx$, where a is the y-intercept and b is the slope, use the following equations:

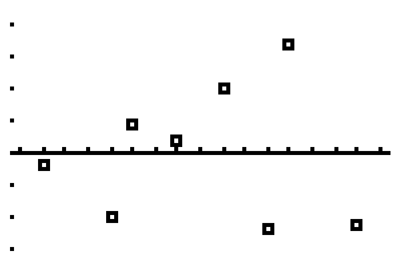
$$b = r\frac{s_y}{s_x} \quad \text{and} \quad a = \bar{y} - b\bar{x}$$

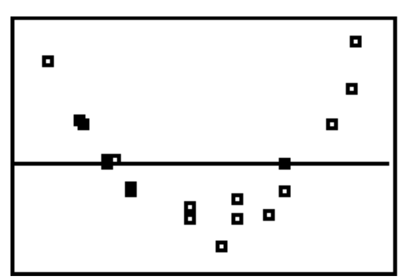
The Role of r-squared (Coefficient of Determination)
The square of the correlation coefficient, or **r-squared**, represents the percentage of the change in the y-variable that can be attributed to its relationship with the x-variable. So if r-squared for the regression between x and y is .73, we can say that x accounts for 73% of the variation in y.

Residuals
A residual is the difference between an observed value of y and the value predicted by the regression line. That is, residual = actual *y* - predicted *y*.

Residual Plot
A residual plot is a scatterplot of each x-value and its residual value. The residual plot is used to determine whether a linear equation is a good model for a set of data, as follows:

- If the residual plot exhibits randomness, then a line is a good model for the data (see left)
- If the residual plot exhibits a pattern, then a line is NOT a good model for the data (right)





Outliers and Influential Points
A point that lies outside the overall pattern of the other observations is considered an **outlier**. If the removal of such a point has a large effect on the correlation and/or regression, that point is considered an **influential point**.