Statistics for the Biological, Environmental and Health Sciences

STAT 007

Correlation and Regression

Chapter 10

Correlation

Section 10-1

- Paired sample data refers to data resulting from the measurement of two variables for each subject. The notation is usually of the form (x, y).
- The graph that we will use to plot paired sample data is called **scatterplot**, which is a cloud of points, each point describing a single sample data pair. On the horizontal axis we plot x, the independent variable, and on the vertical axis we plot y, the dependent variable.
- The numerical measurement of the strength of the *linear* association between two variables is the *linear correlation coefficient, r*, which is a number that measures how well paired sample data fit a straight-line pattern when graphed.
- The linear correlation coefficient, r, is a sample statistic and the population linear correlation coefficient, ρ, is a population parameter.

Correlation

- A **correlation** between two variables exists, when the values of one variable are somehow associated with values of the other variable.
- A **linear correlation** between two variables exists, when there is a correlation and the plotted points of paired sample data result in a pattern that can be approximated by a straight line.
- The **linear correlation coefficient,** r measures the strength of the linear correlation between the paired quantitative x values and y values in a sample.

Properties of Correlation

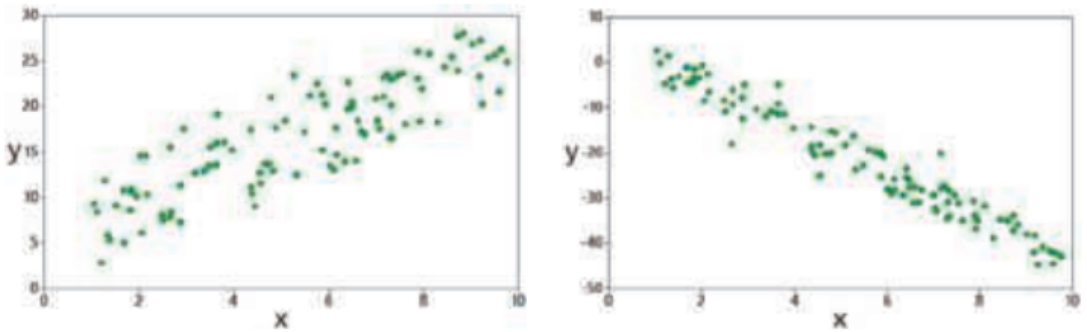
- For n pairs of data values $(x_1, y_1), (x_1 y_2), \ldots, (x_n, y_n)$, the linear correlation coefficient, r is computed as

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}} = \frac{\sum(z_x z_y)}{n-1}.$$

  a) the value of r is always between -1 and 1, this is, $-1 \leq r \leq 1$.
  b) if all values of either variable are converted to a different scale, the value of r does not change.
  c) the value of r is not affected by the choice of x or y. Interchange all x values and y values, and the value of r will not change.
  d) r measures the strength of a linear relationship. It is not designed to measure the strength of a relationship that is not linear.
  e) r is very sensitive to outliers in the sense that a single outlier could dramatically affect its value.

Correlation

- The following are scatterplots together with the computation of the linear correlation coefficient of paired sample data:



(a) Positive correlation: $r = 0.859$

(b) Negative correlation: $r = -0.971$

(c) No correlation: $r = 0.074$

(d) Nonlinear relationship: $r = 0.330$

FIGURE 10-2 Scatterplots

Hypothesis Test for the Population Linear Correlation

- When we want to determine whether there is a significant linear correlation between two variables, perform the following hypothesis test:
  a) $H_0 : \rho = 0 \quad H_1 : \rho \neq 0$.
  b) Set the level of significance, α.
  c) Compute the test statistic $t^{stat} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$, which under the assumption that the null hypothesis is true follows a Student t distribution with $(n-2)$ degrees of freedom.
  d) Conclude to reject when
     — $t^{stat} > t_{\alpha/2}$, if $t^{stat} > 0$
     — $t^{stat} < -t_{\alpha/2}$, if $t^{stat} < 0$

- Requirements to perform a test of hypothesis for ρ:
  a) The sample of paired (x, y) data is a simple random sample.
  b) The points must approximate a straight-line pattern.
  c) Because results can be strongly affected by the presence of outliers, any outliers must be removed if they are known to be errors.
  d) A more formal requirement: The pair of (x, y) data must follow a bivariate normal distribution.

Linear Correlation

Example

The following is paired sample data consisting of pulse rates and white blood cell counts for five adult females.

TABLE 10-1 Pulse Rates and White Blood Cell Counts of Adult Females

Pulse Rate	56.0	82.0	78.0	86.0	88.0
White Blood Cell Count	6.9	8.1	6.4	6.3	10.9

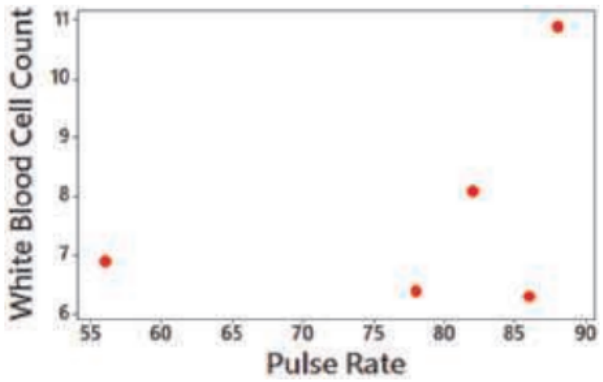


FIGURE 10-1 Scatterplot of Pulse Rates and White Blood Cell Counts

Statistic: $r = 0.405$

Use a 0.05 significance level to test of the claim that there is a linear correlation between the two variables.
What requirements need to be satisfied to perform this test? Write the hypothesis, base your conclusions on the critical value, and discuss the type of error that you could be making with your conclusion.

Common Errors Involving Correlation

- *Assuming that correlation implies causality*.
  One classic example: involves paired data consisting of the stork population in Copenhagen and the number of human births. For several years, the data suggested a linear correlation, but is there a causation?

- *Using data based on averages instead of on individual measurements.* Averages suppress individual variation and may inflate the correlation coefficient.
  One study produced a 0.4 linear correlation coefficient for paired data relating income and education among individuals, but the linear correlation coefficient became 0.7 when regional averages were used.

- Ignoring the possibility of a nonlinear relationship.
  If there is no linear correlation, there might be some other correlation that is not linear.

Practice

Look at the exercises at the end of Section 10-1 in page 456.

Specially, look at exercises: 1, 2, 3, 4, 13, 14, 17, 21, 22. For exercises after 4, you can use technology to compute r and base your conclusions on the critical value.

Regression

Section 10-2

- We will discuss how to find the equation of the straight line that best fits the points in a scatterplot of paired sample data.
- The best fitting line is called the *regression line* and its equations, the *regression equation*.
- The regression equation can be used to make prediction of one of the variables (the dependent variable) given some specific value of the other variable (the independent variable).
- We will also discuss marginal change, influential points, and residual plots to analyze the correlation and regression results.

The Regression Equation

- Given a collection of paired sample data, the **regression line** is the straight line that "best" fits the scatterplot of the data.
- The **regression equation** expresses the relationship between x and $\hat{y}$ as follows

$$\hat{y} = b_0 + b_1 x,$$

where x is called the **independent variable**, and $\hat{y}$ is called the **dependent variable**. b_0 and b_1 are computed as

$$b_1 = r\frac{s_y}{s_x}, \qquad b_0 = \overline{y} - b_1\overline{x},$$

where r is the linear correlation coefficient, s_y is the standard deviation of the y values, and s_x is the standard deviation of the x values.

- b_0 is the y-intercept, b_1 is the slope of the regression equation, and $\hat{y}$ is the predicted value of the dependent variable.
- b_0 is a sample statistic for the population parameter β_0, b_1 is a sample statistic for the population parameter β_1, and $\hat{y} = b_0 + b_1 x$ is a sample statistic for the population parameter $y = \beta_0 + \beta_1 x$.

The Regression Equation

- **Requirements for computing the regression equation**:
  a) The sample of paired (x, y) data is a random sample of quantitative data.
  b) Visual examination of the scatterplot shows that the points approximate a straight-line pattern.
  c) Outliers can have a strong effect on the regression equation, so remove any outliers if they are known to be errors. Consider the effects of any outliers that are not known errors.
  d) For each fixed value of x, the corresponding values of y have a normal distribution.
  e) For the different fixed values of x, the distributions of the corresponding y values all have the same standard deviation.
  f) For the different fixed values of x, the distributions of the corresponding y values have means that lie along the same straight line.

- Requirements b) and c) are informal requirements to check the more formal requirements d), e), and f), which are more difficult to verify.

The Regression Equation

Example

The following is paired sample data consisting of pulse rates and white blood cell counts for five adult females.

TABLE 10-1 Pulse Rates and White Blood Cell Counts of Adult Females

Pulse Rate	56.0	82.0	78.0	86.0	88.0
White Blood Cell Count	6.9	8.1	6.4	6.3	10.9

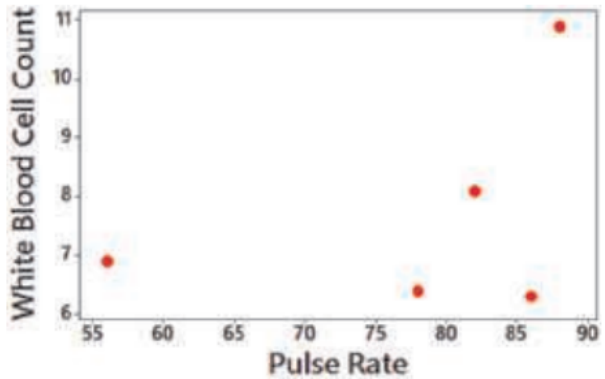


FIGURE 10-1 Scatterplot of Pulse Rates and White Blood Cell Counts

Some statistics:
$r = 0.405, s_y 1.916,$
$s_x = 12.884, \overline{x} = 78,$
$\overline{y} = 7.72$

Find the regression equation in which the independent variable is the pulse rate and the dependent variable is the white blood cell count. Make a plot of your findings.

Making Predictions

- We can use the linear regression equation to make predictions for the dependent variable when:
  a) the graph of the regression line on the scatterplot confirms that the regression line fits the points *reasonably well*.
  b) the linear correlation coefficient r indicates that there is a *linear correlation* between the two variables.
  c) the new value for the independent variable does not go much beyond the scope of the available data.

- If the regression equation does not appear to be useful for making predictions, do not use it!. Better predict the value of the dependent variable as the mean, $\overline{y}$, for every x.

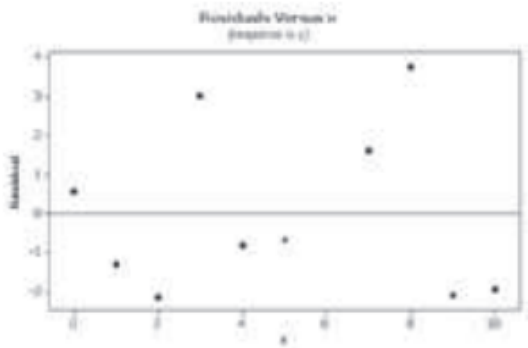
Interpretation, Outliers and Influential Points, and Residuals

- Interpretation: In the regression equation, the slope b_1 represents the marginal change in $\hat{y}$ that occurs when x changes by one unit.
- A correlation/regression analysis should include an investigation of *outliers* and *influential points*:
  a) Outliers are observed in a scatterplot as points that are far away from other data points.
  b) Influential points are points that strongly affect the graph of the regression line.
- A residual plot is a scatterplot of the pairs $(x, y - \hat{y})$, where $y - \hat{y}$ denotes the residuals of the regression model. A regression equation is a good model when the residual plot:
  a) does not have any obvious patterns
  b) does not become much wider (thinner) when viewed from left to right.

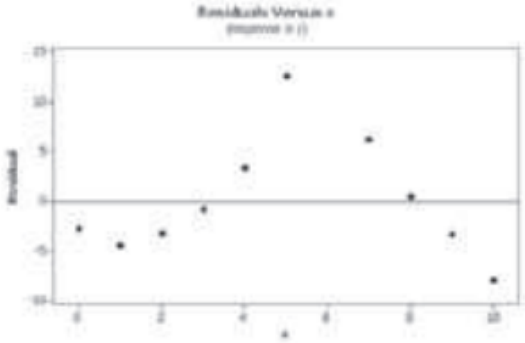
Residuals

- The following plots show residual scatterplots for different regression equations

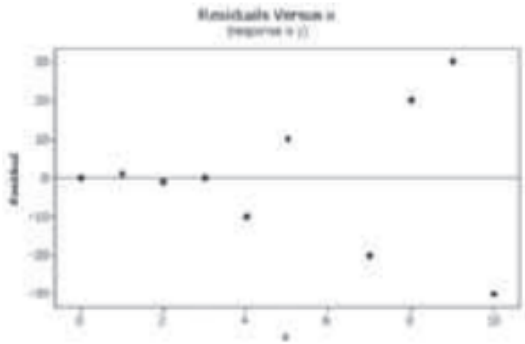
Residual Plot Suggesting That the Regression Equation Is a Good Model



Residual Plot with an Obvious Pattern, Suggesting That the Regression Equation Is Not a Good Model



Residual Plot That Becomes Wider, Suggesting That the Regression Equation Is Not a Good Model

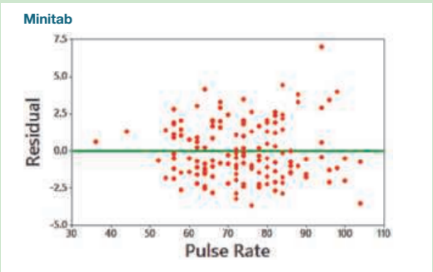


Interpretation, Outliers and Influential Points, and Residuals

Example

From paired sample data consisting of pulse rates and white blood cell counts for 147 adult females, the following statistics were computed: $r = 0.221$, $b_0 = 4.06$, $b_1 = 0.0345$.

a) Determine whether there is a significant linear correlation between pulse rates and white blood cell counts.
b) Write the regression equation with white blood cell counts as the dependent variable and pulse rate as the independent variable.
c) Make an interpretation of b_1.
d) Predict the white blood cell count for a female that has a pulse rate of 80 bpm.
e) The following is a scatterplot of the residuals of this regression equation, does the regression equation be a good model?



Practice

Look at the exercises at the end of Section 10-2 in page 471.

Specially, look at exercises: 1-8, 13, 14, 17, 21, 22. For exercises after 4, you can use technology in your computations.