Correlation

Correlation can be used to investigate the relationship between two continuous variables.

One may also draw a scatterplot of the data (a 2-dimentional plot of one variable vs. the other) to visualize this relationship.
Example: Diphtheria, pertussis, and tetanus (DPT) immunization rates and mortality.

Is there any association between the proportion of newborns immunized and the level of mortality of children less than 5 years of age?
**Scatterplot:** DPT immunization and child mortality

Using visual inspection, we note a general pattern: as the proportion of infants immunized against DPT *increases*, the infant mortality rate *decreases*.
Let:

\[ X = \text{Percent of infants immunized against DPT} \]

\[ Y = \text{Infant mortality (number of infants under 5 dying per 1,000 live births)} \]

A measure of the association between \( X \) and \( Y \) is the *Pearson correlation coefficient* (\( \rho \)), the average of the product of the standardized (normalized) deviates from the mean of each population. It is estimated by

\[
r = \frac{1}{(n-1)} \sum_{i=1}^{n} \left( \frac{x_i - \overline{x}}{s_x} \right) \left( \frac{y_i - \overline{y}}{s_y} \right)
\]

\[
= \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \left[ \sum_{i=1}^{n} (y_i - \overline{y})^2 \right]}}
\]

where \( s_x \) and \( s_y \) are the standard deviations of \( x \) and \( y \) respectively.
The correlation coefficient can take values from -1 to +1.

**Positive** values of $\rho$ imply a *positive* relationship between $x$ and $y$.

**Negative** values of $\rho$ imply a *negative* relationship between $x$ and $y$.

If $|\rho|$ is close to 1, then a *functional* (i.e., deterministic) relationship between $x$ and $y$ exists, meaning that if we know one, it is like knowing the other explicitly.

Independent variables are **uncorrelated**.
Example (continued):

Considering the data from table 17.1, we have the following:

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{20} \sum_{i=1}^{20} x_i = 77.4\%.
\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{1}{20} \sum_{i=1}^{n} y_i = 59.0 \text{ per 1,000 live births}
\]

\[
\sum_{i=1}^{n} (x_i - \bar{x})^2 (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - 77.4)^2 (y_i - 59.0)^2 = 22706
\]

\[
\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{20} (x_i - 0.763)^2 = 10630.8
\quad \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - 0.06222)^2 = 77498
\]
And the correlation coefficient is

\[
r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}
\]

\[
= \frac{-22706}{\sqrt{(10630.8)(77498)}}
\]

\[-0.79\]

Thus there appears to be a negative association between immunization levels for DPT and infant mortality.
Hypothesis testing for $\rho=0$

Tests of hypotheses for zero correlation are based on $se(r) = \sqrt{\frac{1-r^2}{n-2}}$ and the statistic

$$T = \frac{r}{\sqrt{1-r^2}} \cdot \sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2}$$

The test of the null hypothesis of zero correlation is constructed as follows:

1. $H_0$: $\rho=0$
2. a. $H_a$: $\rho>0$  b. $H_a$: $\rho<0$  c. $H_a$: $\rho \neq 0$
3. The level of significance is $(1-\alpha)\%$
4. Rejection rule:
   a. Reject $H_0$ if $t > t_{n-2;1-\alpha}$
   b. Reject $H_0$ if $t < t_{n-2;\alpha}$
   c. Reject $H_0$ if $t > t_{n-2;1-\alpha/2}$ or $t < t_{n-2;\alpha/2}$
Example (continued):

In the previous example, if $\alpha$ is 5% (significance level 95%), since

$$t = r \sqrt{\frac{n-2}{1-r^2}} = -0.79 \sqrt{\frac{20-2}{1-(-0.79)^2}} = -5.47$$

Since $-5.47 < t_{18;0.025}$ we reject the null hypothesis at the 95% level of significance. There seems to be a significant negative correlation between immunization levels and infant mortality. This points to an inversely proportional relationship, i.e., as immunization levels rise, infant mortality decreases.

Note that we cannot estimate how much infant mortality would decrease if a country were to increase its immunization levels by 10%.
### STATA Output

<table>
<thead>
<tr>
<th></th>
<th>immunize</th>
<th>under5</th>
</tr>
</thead>
<tbody>
<tr>
<td>immunize</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>under5</td>
<td>-0.7911</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

**p-value of the test**
CIs for the Correlation Coefficient

The \((1-\alpha)%\) confidence interval for \(\rho\) is based on \(W = \frac{1}{2} \ln \left( \frac{1+r}{1-r} \right)\) where, 

\[ W \sim N \left( 0, \frac{1}{\sqrt{n-3}} \right). \]

The confidence intervals for \(\rho\) are computed in two steps:

**STEP 1.** Compute a confidence interval for the parameter \(W = \frac{1}{2} \ln \left( \frac{1+r}{1-r} \right)\). This is given by the expression 

\[ [L_Z, U_Z] = \left[ \frac{1}{2} \ln \left( \frac{1+r}{1-r} \right) - \frac{z_{1-\alpha/2}}{\sqrt{n-3}} \ln \left( \frac{1+r}{1-r} \right) + \frac{z_{1-\alpha/2}}{\sqrt{n-3}} \right]. \]

**STEP 2.** Solve equations 

\[ L_Z = \frac{1}{2} \ln \left( \frac{1+L_\rho}{1-L_\rho} \right) \iff \frac{e^{2L_Z} - 1}{e^{2L_Z} + 1} \] and 

\[ L_Z = \frac{1}{2} \ln \left( \frac{1+U_\rho}{1-U_\rho} \right) \iff \frac{e^{2U_Z} - 1}{e^{2U_Z} + 1}. \]

Thus, a \((1-\alpha)%\) confidence interval for \(\rho\) is given by

\[
\left[ \frac{e^{2L_Z} - 1}{e^{2L_Z} + 1}, \frac{e^{2U_Z} - 1}{e^{2U_Z} + 1} \right].
\]
Example: In this example, the 95\% confidence interval is as follows:

**STEP 1.** A confidence interval for \( W \) is as follows:

\[
\left[ L_z, U_z \right] = \left[ \frac{1}{2} \ln \left( \frac{1+r}{1-r} \right) - \frac{z_{1-\alpha/2}}{\sqrt{n-3}}, \ \frac{1}{2} \ln \left( \frac{1+r}{1-r} \right) + \frac{z_{1-\alpha/2}}{\sqrt{n-3}} \right]
\]

\[
= \left[ \frac{1}{2} \ln \left( \frac{1+(-0.79)}{1-(-0.79)} \right) - \frac{1.96}{\sqrt{17}}, \ \frac{1}{2} \ln \left( \frac{1+(-0.79)}{1-(-0.79)} \right) + \frac{1.96}{\sqrt{17}} \right] = (-1.547, -0.596)
\]

**STEP 2.** The confidence interval for \( \rho \) is then,

\[
\left[ \frac{e^{2L_z} - 1}{e^{2L_b} + 1}, \frac{e^{2U_z} - 1}{e^{2U_b} + 1} \right] = \left[ \frac{e^{(2.0)(-1.547)} - 1}{e^{(2.0)(-1.547)} + 1}, \frac{e^{(2.0)(-0.596)} - 1}{e^{(2.0)(-0.596)} + 1} \right] = (-0.913, -0.534)
\]

This interval does not contain zero (the hypothesized value under the null hypothesis in the previous test). This is consistent to rejecting the hypothesis of zero correlation between immunization levels and infant mortality. This is not always the case as the hypothesis test and the confidence interval are based on different statistics.
Simple Linear Regression

There is a mathematical way characterize a straight line. We specify the relationship between two continuous variables $x$ (measured on the horizontal or $x$ axis) and $y$ (measured on the vertical or $y$ axis).

The linear equation relating the $x_i$ to the $y_i$ is as follows:

$$y_i = \beta_0 + \beta_1 x_i$$

$\beta_0$ is called the intercept of the line (because if $x_i=0$ the line “intercepts” the $y$ axis at $\beta_0$), and $\beta_1$ is called the slope of the line.
I. Both lines have the same intercept but different positive slopes.

II. Both lines have the same positive slope (they are parallel) but different intercepts.

III. Both lines have the same intercept but different negative slopes

IV. Both lines have the same negative slope (parallel) but different intercepts.
One attractive feature of a linear relationship is the *constant slope*. This means that for a fixed increase $\Delta x$ in $x$, there will be a fixed change $\Delta y$ ($= \beta_1 \Delta x$) regardless of the value of $x$ (e.g., there is a fixed *increase* if the slope is positive and a fixed *decrease* if the slope is negative). This is in contrast to a *non-linear* relationship, such a *quadratic* or *polynomial*, where for some values of $x$, $y$ will be increasing, and for some other values $y$ will be decreasing (or vice versa).
Consider the following scatterplot

![Scatterplot](image)

Even though it seems that $y$ may be increasing for increasing $x$, the relationship is not a perfect line. If we want to draw a line through the observations that we think best describes the trends in our data we may be confronted with many candidate lines.
Determining the best regression line

Consider the following figure:

\[ \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \]
The least-squares method

The regression line (whichever it is) will not pass through all data points \( Y_i \). Thus, in most cases, for each point \( X_i \) the line will produce an estimated point \( \hat{Y}_i = \beta_0 + \beta_1 X_i \) and most probably, \( \hat{Y}_i \neq Y_i \). In fact, as we see in the previous figure, \( Y_i = \hat{Y}_i + e_i \). For each choice of \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) (note that each pair \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) completely defines the line) we get a new line, and a whole new set of deviation terms \( e_i \).

The “best-fitting line” according to the least-squares method is the one that

\[
\text{minimizes} \quad \sum_{i=1}^{n} \left( Y_i - \hat{Y}_i \right)^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[ Y_i - \left( \beta_0 + \beta_1 X_i \right) \right]^2.
\]
Explaining variability

The total variability in the data is given by

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}[ (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})]^2$$

as we can see by inspection of the previous figure. It turns out that

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

unexplained variability

variability due to regression

$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ is the part of the total variability that can be explained while

$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ is left unexplained, because the model cannot account for differences between the estimated points and the data (this is called error sum of squares).
Degrees of freedom

1. $S_Y^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \frac{1}{(n-1)} SSY$
2. $S_{Y|X}^2 = \frac{1}{(n-2)} \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = \frac{1}{(n-2)} SSE$

$S_Y^2$ is the variability of $Y$ without consideration of $X$ and $S_{Y|X}^2$ is the residual variability after the relationship of $Y$ and $X$ has been accounted for.

The total sum of squares, $SSY$, is made up of $n$ terms $(Y_i - \bar{Y})^2$. Once the mean $\bar{Y}$ has been estimated, only $n-1$ terms are needed to compute $SSY$. The degrees of freedom of $SSY$ are $n-1$. The sum of squares due to regression, $SSR$ is computed from a single function of $Y$, that is, $SSR = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \beta_1 \sum_{i=1}^{n} (X_i - \bar{X})^2$ and has only one degree of freedom. $SSE$ has the remaining $n-2$ degrees of freedom.
The Analysis of Variance Table

<table>
<thead>
<tr>
<th>Source of variability</th>
<th>Sums of squares (SS)</th>
<th>Df</th>
<th>Mean squares (MS)</th>
<th>F</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>$SSR = \hat{\beta}<em>1 \sum</em>{i=1}^{n}(X_i - \bar{X})^2$</td>
<td>1</td>
<td>$MSR = SSR$</td>
<td>$F = \frac{MSR}{MSE}$</td>
<td>$P = P(F &gt; F_{1, n-2; \alpha})$</td>
</tr>
<tr>
<td>Residual (error)</td>
<td>$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$</td>
<td>$n-2$</td>
<td>$MSE = \frac{SSE}{n-2}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>$SSY = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$</td>
<td>$n-1$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Assumptions of the linear regression model

1. The $y$ values are distributed according to a normal distribution with mean $\mu_{y|x}$ and variance $\sigma_{y|x}$ that is unknown.

2. The relationship between $X$ and $Y$ is given by the formula $\mu_{y|x}=\beta_0+\beta_1X$.

3. The $y$ are independent.

4. For every value $x$ the standard deviation of the outcomes $y$ is constant and equal to $\sigma_{y|x}$. This concept is called homoscedasticity.
Tests of hypothesis

In these models, $y$ is our target (or dependent variable, the outcome that we cannot control but want to explain) and $x$ is the explanatory (or independent variable).

Within each regression, the primary interest is the assessment of the existence of the linear relationship between $x$ and $y$. If such an association exists, then $x$ provides information about $y$.

Inference on the existence of the linear association is accomplished via tests of hypotheses and confidence intervals. Both of these center around the estimate of the slope $\hat{\beta}_1$, since it is clear, that if the slope is zero, then changing $x$ will have no impact on $y$ (thus there is no association between $x$ and $y$).
The $F$ test of overall linear association

The test of hypothesis of no linear association is defined in a similar manner as in the ANOVA:

1. $H_0$: No linear association exists between $x$ and $y$.
2. $H_a$: A linear association exists between $x$ and $y$.
3. Tests are carried out at the $(1-\alpha)\%$ level of significance
4. The test statistic is $F = \frac{MSR}{MSE}$, where the numerator is akin to the *between groups* mean squares, and the denominator is similar to the *within groups* mean squares.
5. **Rejection rule:** Reject $H_0$, if $F > F_{1, n-2; \alpha}$. This will happen if $F$ is far from unity (just like in the ANOVA case).
The $F$ test (continued)

The $F$ test of linear association tests whether a line, other than the horizontal one going through the sample mean of the $Y$’s, is useful in explaining some of the variability in the data. It is based on the fact that, $MSR \approx \sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2$ for large $n$, while $MSE \approx \sigma^2$. If the population regression slope $\beta_1 \approx 0$, that is, if the regression does not add anything new to our understanding of the data (does not explain a substantial part of the variability), then the two mean square errors $MSR$ and $MSE$ are estimating a common quantity (the population variance $\sigma^2$).

Thus the ratio should be close to 1 if the hypothesis of no linear association between $X$ and $Y$ is present. On the other hand, if a linear relationship exists, ($\beta_1$ is far from zero) then $SSR > SSE$ and the ratio will deviate significantly from 1.
Hypothesis testing for zero slope

The test of hypothesis of no linear association is defined as follows:

1. $H_0$: No linear association between $x$ and $y$: $\beta_1=0$.

2. $H_a$: A linear association exists between $x$ and $y$:
   a. $\beta_1 \neq 0$ (two-sided test)
   b. $\beta_1 > 0$
   c. $\beta_1 < 0$ (one-sided tests)

3. Tests are carried out at the $(1-\alpha)$% level of significance

4. The statistic $T = \frac{\hat{\beta}_1}{\text{s.e.}(\hat{\beta}_1)}$ is distributed as a $t$ distribution with $n-2$ d.f.

5. Rejection rule: Reject $H_0$, in favor of the three alternatives respectively, if
   a. $t < t_{n-2;\alpha/2}$, or $t > t_{n-2;(1-\alpha/2)}$
   b. $t > t_{n-2;(1-\alpha)}$
   c. $t < t_{n-2;\alpha}$
**CIs for \( \beta_1 \)**

Confidence intervals of \( \beta_1 \) are constructed as usual and are based on \( \hat{\beta}_1 \), the standard error of \( \hat{\beta}_1 \), and the \( t \) statistic.

A \((1-\alpha)\)% confidence interval is as follows:

\[
\left[ \hat{\beta}_1 - t_{n-2; (1-\alpha/2)} \text{s.e.}(\hat{\beta}_1), \hat{\beta}_1 + t_{n-2; (1-\alpha/2)} \text{s.e.}(\hat{\beta}_1) \right]
\]
Example: Newborn Infant Length and Gestational Age

Consider the problem described in section 18.4 of the textbook. Is there a relationship between the length and the gestational age of newborn infants?

Scatterplot of length vs. gestational age:
Example (continued): STATA Output

```
<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>575.73916</td>
<td>1</td>
<td>575.73916</td>
<td>F( 1, 98) = 82.13</td>
</tr>
<tr>
<td>Residual</td>
<td>687.02084</td>
<td>98</td>
<td>7.01041674</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>1262.76</td>
<td>99</td>
<td>12.7551515</td>
<td>R-squared = 0.4559</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adj R-squared = 0.4504</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Root MSE = 2.6477</td>
</tr>
</tbody>
</table>

length       | Coef.  | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
-------------|--------|-----------|-------|------|---------------------|
gestage      | .9516035 | .1050062 | 9.062 | 0.000 | .7432221   1.159985 |
_cons        | 9.3281730 | 3.0451630 | 3.063 | 0.003 | 3.285149  15.3712  |
```

A: Source | SS    | df | MS     |
B: Model | 575.73916 | 1  | 575.73916 |
C: Residual | 687.02084  | 98 | 7.01041674 |
D: Total | 1262.76 | 99 | 12.7551515 |
E: length | Coef.  | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
F: gestage | .9516035 | .1050062 | 9.062 | 0.000 | .7432221   1.159985 |
G: _cons   | 9.3281730 | 3.0451630 | 3.063 | 0.003 | 3.285149  15.3712  |
Computer Output

A. Degrees of freedom. There is one degree of freedom associated with the model, and \( n-2=98 \) degrees of freedom comprising the residual.

B. \( F \) test. This is the overall test of linear association. Note that the numerator degrees of freedom is the model degrees of freedom, while the denominator degrees of freedom is the residual (error) degrees of freedom. This test measures deviations from one.

C. Rejection rule of the \( F \) test. Since the p-value is \( 0.000<0.05=\alpha \), we reject the null hypothesis. There is evidence to suggest a linear association between gestational age and length of the newborn.

D. Root MSE. This is the square root of the mean squares for error, and as before, can be used as an estimate of the population variability.
Computer implementation (continued):

A. `gestage` is the estimate of the slope, $b=0.9516035$. This means that the fetus grows by 0.95 inches every week of gestation.

B. **p-value** of the $t$ test described above. Since $0.000<0.05$, we reject the null hypothesis. There is a strong positive linear relationship between gestational age and length of the newborn.

C. The 95% confidence interval for $\beta$ is $[0.743, 1.160]$. Since it excludes 0, we reject the null hypothesis and conclude that there is a strong positive linear relationship between gestational age and length of newborn.

D. The estimate of the intercept `_cons`=9.328 implies that the fetus is 9.3 inches long at conception (week zero) which is totally false. Note however, that our model is only valid in the period in which our data have been collected.
What the correlation does not measure

1. The correlation coefficient does not measure the magnitude of the regression slope. Even though, \( r_{xy} = \frac{S_x}{S_y} \hat{\beta}_1 \Leftrightarrow \hat{\beta}_1 = \frac{S_y}{S_x} r_{xy} \), the size of the slope is also dependent on the variability in the \( X \) and \( Y \)

2. The correlation coefficient is not a measure of the appropriateness of the straight-line model. Consider the case of a quadratic model ("U"-shaped curve; see figures II and IV below). The correlation coefficient may be large, but a straight-line model may still be inappropriate (figure IV).
Final remark. Beyond the limitations of the correlation coefficient just mentioned, there are additional advantages in using a regression model. Consider the DPT immunization example:

![Regression Table](image-url)
Final remark (continued)

The estimates of slope and intercept are $\hat{\beta}_0 = 224.32$ and $\hat{\beta}_1 = -2.136$ respectively.

If an official wanted to estimate the decline in infant mortality for a 10% increase in immunization from current levels they would only have to multiply 10 (=10%) by the estimate for the slope $-2.136$ (taking advantage of the fact that for a constant increase in $x$, $y$ decreases by a constant factor).

If DPT immunization rates were increased by 10%, the infant mortality would drop by $(10)(-2.136) = -21.36$ or about 21 infants per 1,000 live births. This would account for one-third the mortality rate of Egypt, or two-thirds the infant mortality rate in Russia. Taking advantage of the properties of a straight line and the results of the regression model, a concrete public health policy could be constructed.