Chapter 12.5: Correlation

Instructor: Dr. Arnab Maity
Suppose we have two variables $x$ and $y$. Regression analysis uses one variable ($x$) to predict the other ($y$) using a linear relationship. However, often we may want to look at joint behavior of the two variables together rather than predicting one using the other. In this chapter, we learn about the sample correlation coefficient $r$ as a measure of how strongly two variables $x$ and $y$ are related in a sample.

**Recall:** Given two random variables $X$ and $Y$, the correlation coefficient or the population correlation coefficient is defined as

$$
\rho = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}.
$$

This measures the strength of relationship between $X$ and $Y$.

The correlation coefficient has the following properties.

- It does not have any units.
- $-1 \leq \rho \leq 1$.
- If $X$ and $Y$ are independent, then $\rho = 0$, but $\rho = 0$ does not imply independence.
- $\rho = -1$ or $\rho = 1$ if and only if $Y = a + bX$ for some number $a$ and nonzero number $b$ (that is, $Y$ and $X$ are linearly related).

These properties tell us that $\rho$ measures the strength of linear relationship between $X$ and $Y$. When $\rho = 0$, $X$ and $Y$ are said to be uncorrelated. They can still be dependent if there is strong nonlinear relationship between them.

**Example:** Let $X$ be a discrete random variable which takes three values -1, 0 and 1 with probability 1/3 for each. Let $Y = X^2$. Then we have

$$
E(X) = 0, E(XY) = 0
$$

and as a result $Cov(X,Y) = 0$. This also means $\rho = 0$. But as we clearly see $X$ and $Y$ are clearly dependent (as $Y = X^2$).

A value of $\rho$ near 1 does not necessarily imply that increasing the value of $X$ causes $Y$ to increase. It implies only that large $X$ values are associated with large $Y$ values. Be careful: association (a high correlation) is not the same as causation.
Now we will define a sample version of the correlation coefficient. Suppose that we observe \( n \) pairs of data points \((x_1, y_1), \ldots, (x_n, y_n)\). It is natural to say that the variables \( x \) and \( y \) are positively related if large values of \( x \) are paired with large values of \( y \) and small values of \( x \) are paired with small values of \( y \). Similarly, we may say that \( x \) and \( y \) are negatively related if large values of \( x \) are paired with small values of \( y \) and small values of \( x \) are paired with large values of \( y \).

To define the sample version of correlation, we first define the quantity

\[
S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).
\]

Note that the two variables are positively correlated, then \( x_i \) values above \( \bar{x} \) will tend be paired with \( y_i \) values above \( \bar{y} \). Thus the product \((x_i - \bar{x})(y_i - \bar{y})\) will be positive. Similarly, \( x_i \) values below \( \bar{x} \) will tend be paired with \( y_i \) values below \( \bar{y} \) again making the product \((x_i - \bar{x})(y_i - \bar{y})\) positive. Thus the quantity \( S_{xy} \) will be positive if there is a positive relation between \( x \) and \( y \). A similar argument can be made for negative relationship as well.

The **sample correlation coefficient** is defined as

\[
r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}},
\]

where

\[
S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad \text{and} \quad S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2.
\]
Properties of $r$

1. The value of $r$ does not depend on which of the two variables under study is labeled $x$ and which is labeled $y$.

2. The value of $r$ is independent of the units in which $x$ and $y$ are measured.

3. $-1 \leq r \leq 1$

4. $r = 1$ if and only if all $(x_i, y_i)$ pairs lie on a straight line with positive slope, and $r = -1$ if and only if all $(x_i, y_i)$ pairs lie on a straight line with negative slope.

5. The square of the sample correlation coefficient gives the value of the coefficient of determination that would result from fitting the simple linear regression model.

Property 1 is exactly opposite to what happens in regression analysis, where all the quantities of interest (e.g., slope, intercept and error variance) depend on which of the two variables is the response/dependent variable.

On the other hand, Property 5 tells us that the proportion of variation in the dependent variable explained by fitting the simple linear regression model does not depend on which of the two variables is the response/dependent variable.

**Example:** Recall the Arsenic data example discussed in the previous lecture. We saw data on $x =$ pH and $y =$ arsenic removed (%) by a particular process. Data for this example is shown in the book (Example 12.2).

Simple linear regression results:
Dependent Variable: Arsenic_removed  
Independent Variable: pH  
Arsenic_removed = 190.268 - 18.034 pH  
Sample size: 18  
$R$ (correlation coefficient) = -0.950  
$R$-sq = 0.903  
Estimate of error standard deviation: 6.125

Simple linear regression results:
Dependent Variable: pH  
Independent Variable: Arsenic_removed  
pH = 10.350 - 0.050 Arsenic_removed  
Sample size: 18  
$R$ (correlation coefficient) = -0.950  
$R$-sq = 0.903  
Estimate of error standard deviation: 0.322
Property 2 indicates that the sample correlation coefficient remains the same even we change the scale of measurements (e.g., meters vs. feet, kilograms vs. pounds) or a location shift.

Property 3 tells us that \( r \) can only take value between -1 and +1. This is exactly the same as that of the population correlation coefficient \( \rho \). Combined with property 4, we can see that \( r = 1 \) if and only if all the \( x \) and \( y \) pairs fall exactly on a straight line with positive slope. A similar statement can be made for \( r = -1 \).

\( r \) will be less than 1 in absolute value for any other type of scatterplot even if there are clear relationship between \( x \) and \( y \).

An informal rule of thumb for the value of \( r \) is given below.

- Weak: \(-0.5 \leq r \leq 0.5 \) \((R^2 \leq 0.25)\)
- Moderate: \(-0.8 < r < 0.5 \) or \( 0.5 < r < 0.8 \) \((0.25 < R^2 < 0.64)\)
- Strong: \( r \leq -0.8 \) or \( r \geq 0.8 \) \((R^2 \geq 0.64)\)
The correlation coefficient $r$ measures how strong the linear relationship between $x$ and $y$ is in the observed sample. We can think of the pairs $(x_i, y_i)$ as random samples from a population with unknown population correlation coefficient $\rho$. In fact, we can think the sample correlation coefficient $r$ as a point estimate of the population correlation coefficient $\rho$ based on the observed sample. This lets us to do inference on $\rho$.

Assumptions:
- $X_1, \ldots, X_n$ are random sample from a normal distribution $N(\mu_X, \sigma^2_X)$
- $Y_1, \ldots, Y_n$ are random sample from a normal distribution $N(\mu_Y, \sigma^2_Y)$
- Correlation coefficient between $X$ and $Y$ is $\rho$.

Define the point estimator for $\rho$ as

$$R = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2}}.$$

Hence, the sample correlation coefficient $r$ is a point estimate of $\rho$.

---

**Testing for the Absence of Correlation**

1. We want to test $H_0 : \rho = 0$
2. Test statistic:

$$T = \frac{R\sqrt{n - 2}}{\sqrt{1 - R^2}}$$

3. If the null hypothesis $H_0$ is true, $T$ follows a $t$-distribution with $n - 2$ degrees of freedom.
4. Value of the test statistic:

$$t = \frac{r\sqrt{n - 2}}{\sqrt{1 - r^2}}$$

5. Rejection regions:

<table>
<thead>
<tr>
<th>Alternative Hypothesis</th>
<th>Rejection Region for a Level $\alpha$ Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_a : \rho \neq 0$</td>
<td>$t \leq -t_{\alpha/2,n-2}$ OR $t \geq t_{\alpha/2,n-2}$ (two-tailed)</td>
</tr>
<tr>
<td>$H_a : \rho &lt; 0$</td>
<td>$t \leq -t_{\alpha,n-2}$ (lower-tailed)</td>
</tr>
<tr>
<td>$H_a : \rho &gt; 0$</td>
<td>$t \geq t_{\alpha,n-2}$ (upper-tailed)</td>
</tr>
</tbody>
</table>

6. **Caution**: The assumptions of normality of $X$ and $Y$ are crucial to employ this test. If the data deviates from normality substantially (e.g., a Q-Q plot shows deviation from straight line), this test should not be employed for small sample sizes.
**Example:** We revisit the arsenic example.

Simple linear regression results:
Dependent Variable: Arsenic_removed
Independent Variable: pH
Arsenic_removed = 190.26829 - 18.034245 pH
Sample size: 18
R (correlation coefficient) = -0.95049529
R-sq = 0.9034413
Estimate of error standard deviation: 6.1255839

Parameter estimates:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Err.</th>
<th>Alternative</th>
<th>DF</th>
<th>T-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>190.26829</td>
<td>12.587118</td>
<td>≠ 0</td>
<td>16</td>
<td>15.116112</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Slope</td>
<td>-18.03425</td>
<td>1.4739533</td>
<td>≠ 0</td>
<td>16</td>
<td>-12.23529</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

Test $H_0: \rho = 0$ vs. $H_a: \rho \neq 0$.

**Relationship with model utility test in regression analysis:** they are equivalent. (compare with test statistic for $\beta_1$)
Other inferences for $\rho$

1. Under the assumptions made before, the quantity

$$V = \frac{1}{2} \ln \left( \frac{1 + R}{1 - R} \right)$$

has \textit{approximately} a normal distribution with mean and variance

$$\mu_V = \frac{1}{2} \ln \left( \frac{1 + \rho}{1 - \rho} \right) \quad \text{and} \quad \sigma_V^2 = \frac{1}{n - 3}.$$

2. We want to test $H_0 : \rho = \rho_0$.
   Here $\rho_0$ can be any value, not just zero.

3. Test statistic:

$$T = \sqrt{n - 3} \left[ V - \frac{1}{2} \ln \left( \frac{1 + \rho_0}{1 - \rho_0} \right) \right].$$

4. For \textbf{large sample size} $n$, if the null hypothesis $H_0$ is true, $Z$ follows $N(0, 1)$ distribution.

5. Rejection regions:

<table>
<thead>
<tr>
<th>Alternative Hypothesis</th>
<th>Rejection Region for a Level $\alpha$ Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_a : \rho \neq 0$</td>
<td>$z \leq -z_{\alpha/2}$ OR $z \geq z_{\alpha/2}$ (two-tailed)</td>
</tr>
<tr>
<td>$H_a : \rho &lt; 0$</td>
<td>$z \leq -z_\alpha$ (lower-tailed)</td>
</tr>
<tr>
<td>$H_a : \rho &gt; 0$</td>
<td>$z \geq z_\alpha$ (upper-tailed)</td>
</tr>
</tbody>
</table>

6. \textbf{Caution}: This test should not be employed for small sample sizes.