Simple Linear Regression II – Test of Significance

How good is our correlation calculation, i.e., when would we be able to conclude there is a significant linear correlation? How good are the predictions based on that conclusion?

We ask the following conditions to hold:

(i) the data is from simple random sample (sample points are independent)

(ii) for any given value (instance) of $X$ variable we have $Y$ distributed normally with the same normal distribution. The same should hold when we switch variables.

If these conditions hold then the variable

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

is a Student’s $t$ random variable with $n - 2$ degrees of freedom (where $R$ is a correlation coefficient of a sample of $n$ pairs of data) in case when $X$ and $Y$ are independent. The instance of this variable

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

(1.1)

calculated from our sample at hand will be the test statistic for our test of linear correlation.

The test is usually a two-tailed test (it is possible to conduct one-tailed test as well but it is uncommon). Both hypothesis are made in regard to population correlation coefficient $\rho$. This is the correlation coefficient we would calculate from the whole population in the same manner we calculate $r$ from any sample. The null hypothesis is that there is no correlation i.e.,

$$H_0 : \rho = 0$$

And the alternative is that there is a correlation

$$H_1 : \rho \neq 0.$$
The critical values to consider will be two-tailed critical values for \( t \) distribution. Rejection of the null hypothesis would indicate there is a linear correlation between \( X \) and \( Y \).

Example 1.

The following data is collected for 8 students attending the same calculus class. We are examining the relation between number of absences and the final test score for these students. The \( X \) variable is the number of absences, the \( Y \) is the respective final test score.

<table>
<thead>
<tr>
<th>( x )</th>
<th>( y )</th>
<th>( xy )</th>
<th>( x^2 )</th>
<th>( y^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>77</td>
<td>77</td>
<td>1</td>
<td>5929</td>
</tr>
<tr>
<td>2</td>
<td>65</td>
<td>130</td>
<td>4</td>
<td>4225</td>
</tr>
<tr>
<td>0</td>
<td>59</td>
<td>0</td>
<td>0</td>
<td>3481</td>
</tr>
<tr>
<td>4</td>
<td>67</td>
<td>268</td>
<td>16</td>
<td>4489</td>
</tr>
<tr>
<td>0</td>
<td>96</td>
<td>0</td>
<td>0</td>
<td>9216</td>
</tr>
<tr>
<td>1</td>
<td>82</td>
<td>82</td>
<td>1</td>
<td>6724</td>
</tr>
<tr>
<td>2</td>
<td>51</td>
<td>102</td>
<td>4</td>
<td>2601</td>
</tr>
<tr>
<td>2</td>
<td>68</td>
<td>136</td>
<td>4</td>
<td>4624</td>
</tr>
<tr>
<td>12</td>
<td>565</td>
<td>795</td>
<td>30</td>
<td>41289</td>
</tr>
</tbody>
</table>

We are not surprised by \( r \) being negative since that means “the more absences the lower score.” However we did not expect the correlation to be so “weak.” Let’s test this correlation using 5% significance.

\[
H_0 : \rho = 0 \\
H_1 : \rho \neq 0
\]

\[
T : r \frac{n-2}{\sqrt{1-r^2}} \approx -0.4071 \sqrt{\frac{8-2}{1-0.1657}} \approx -1.092
\]

The next we need critical values for the \( t \) distribution with 6 degrees of freedom

\[
t_{0.025} = 2.447, -t_{0.025} = -2.447
\]

See the figure 1 below.
The test statistic is well within confidence region thus we should not reject the null hypothesis. That means that the correlation is not significant enough to use simple regression for prediction. It does not mean that there is no correlation as that would mean that we are accepting null hypothesis and we shall not do so.

Example 2.

In our quiz-quiz example (see lecture #11) we used least-squares regression line to make predictions. Was that justified? We can set up the test to determine this at 5% significance.

\[
\begin{align*}
H_0 : \rho &= 0 \\
H_1 : \rho &\neq 0 \\
T : r \sqrt{\frac{n - 2}{1 - r^2}} &\approx 0.892 \sqrt{\frac{10 - 2}{1 - 0.796}} \approx 5.586
\end{align*}
\]

Checking critical values for 8 degrees of freedom gives \( t_{0.025} = 2.306, -t_{0.025} = -2.306 \). See the figure 2 below. We can reject the null hypothesis and claim the significant correlation. Therefore using least-squares regression line for predictions is justified.
Note on One-tail Tests

In the previous example it would be appropriate to conduct right-tail test since there is an obvious indication from the sample data. This means that we suspect there is a positive correlation $H_1 : \rho > 0$.

Thus the setup of the test would be:

$$H_0 : \rho = 0$$
$$H_1 : \rho > 0$$

$$T : r \sqrt{\frac{n-2}{1-r^2}} \approx 0.892 \sqrt{\frac{10-2}{1-0.796}} \approx 5.586$$

Now there is a change of critical value since we are looking only for a rejection region in the right tail. Thus we have $t_{0.05} = 1.860$ for the critical value and yet again because the test value is more extreme than the critical value we would reject null hypothesis and we would claim that there is a significant positive correlation (which is the alternative in this test). See figure 3 below.
In the same manner we could conduct left-tail test in the example #1 using alternative hypothesis as $H_1: \rho < 0$ i.e., negative correlation. [Done in class.]

Note on “Coefficient of Determination”

We use a square of $r$ i.e., $r^2$ to claim the percentage of variation of response variable ($Y$) related to variation of explanatory variable ($X$).

Homework

Review Homeworks 10.1 and 10.2 and test the significance of correlation coefficient in the problems from that homework. Use level of significance 5% in all the tests.