MIOT H5014

Statistical Analysis for Engineers

Worksheet 5 on Hypothesis Testing

Semester 2, 2016/7

Question 1

A lecturer is seeking to prove that attendance is a strong predictor of final marks for a group of students. For 14 students, the attendance over the course of the semester and the final mark were recorded. The correlation between the two variables will be investigated with an NHBST.

1. It is proposed to test the Null Hypothesis that $\rho = 0$ with a transformation of r to a variable that follows the Student t-distribution. Before gathering any data, set up the Null and alternative Hypothesis and other elements of the test and state the assumptions made about the model and data.

Solution

The Null hypothesis: $\rho = 0$.

The Alternative hypothesis: we are assuming that the more a student comes in, the better they do so the alternative hypothesis will then be that there is a positive correlation between attendance and final mark; $\rho > 0$. This means we are doing a one-tailed test.

Use a level of significance 0.05.

The test assumes that the two variables of attendance and final mark are a bivariate normal distribution.

Since the variable r is transformed into a t-distribution variable, we use the t-tables for the critical values at $n - 2$ degrees of freedom.

The critical value for the t distribution at 12 degrees of freedom for 0.05 is 1.782. For 0.01 it is 2.681

2. The data has been collected and the values are given in the table shown. Calculate r and carry out the test.

Solution

Calculate r:

$$r = \frac{\sum X_i Y_i - n\overline{X}\,\overline{Y}}{\sqrt{\sum X_i^2 - n\overline{X}^2}\sqrt{\sum Y_i^2 - n\overline{Y}^2}}.$$

This works out to be:

$$r = \frac{48{,}082 - 14 \times 57.5 \times 56.2143}{\sqrt{51{,}841 - 14 \times 57.5^2}\sqrt{48{,}082 - 14 \times 56.2143^2}} =$$

$$\frac{2829.5}{\sqrt{5{,}553.5}\sqrt{3{,}170.357}} = 0.674.$$

To test this value, calculate t:

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0.674\sqrt{\frac{12}{1-0.674^2}} = 3.163.$$

This is very much greater than the critical value, including at a significance level of 0.01, so therefore we reject the null hypothesis; there is very strong evidence of a positive correlation between the attendance and the final mark.

3. Repeat the two previous steps with the Fisher transformation of ρ.

Solution

With the Fisher transformation we need not start with the Null Hypothesis that $\rho = 0$. Given that we are interested in a positive correlation, we could start with a value we might think is a good one, e.g. $\rho = 0.5$. We will do both.

The test statistic follows the standard normal distribution, so we will look for 1.65 for l.o.s of 0.05 or 2.33 for 0.01.

The test proceeds as before:

$H_0 : \rho = 0$, $H_A : \rho > 0$ (we are investigating a positive correlation)

This means we are doing a one tailed test.

The calculations are shown here.

With $r = 0.674$, then $(1 + r)/(1 – r) = 5.1350$ and so

$F(r) = ½ \log_e(5.1350) = 0.8181$.

With $\rho = 0$, then $(1 + \rho)/(1 – \rho) = 1$ and $F(\rho) = 0$.

Then our value of the test statistic is $[F(r) – F(\rho)]\sqrt{11} = 2.713$.

This is above the critical value for either l.o.s so we reject the Null Hypothesis.

Alternatively:

$H_0 : \rho = 0.5$, $H_A : \rho > 0.5$ (we are investigating a higher correlation)

This means we are doing a one tailed test.

The calculations are shown here.

With $r = 0.674$, then $(1 + r)/(1 – r) = 5.1350$ and so

$F(r) = ½ \log_e(0.1947) = 0.8181$.

With $\rho = 0.5$, then $(1 + \rho)/(1 – \rho) = 3$, so $F(\rho) = ½ \log_e 3 = 0.5493$.

Then our value of the test statistic is $[F(r) – F(\rho)]\sqrt{11} = 0.8915$.

This is nowhere near the critical value for the 0.05 l.o.s so we do not reject the Null Hypothesis.

4. Use the inverse of the Fisher transformation to calculate a confidence interval for ρ.

Solution

The limits of the confidence interval for $F(\rho)$ are given by

$$F(r) \pm \frac{z_{\alpha/2}}{\sqrt{n-3}}.$$

Recall that $F(r) = ½ \log_e(5.1350) = 0.8181$.

Then the limits for 0.05 are $0.8181 \pm 1.96/\sqrt{11}$.

These two numbers are 0.2271 and 1.4090.

Converting them back into correlation values using $F^{-1}(x) = \tanh(x)$, this is 0.2233 to 0.8873.

While this may seem a wide range, importantly both limits are on the positive side. The range is wide since the sample size is small.

Data:

Att.:	29	84	91	78	24	42	65
Mark:	47	73	76	33	32	57	64
Att.:	57	38	58	48	69	76	46
Mark:	54	38	65	50	75	75	48

Question 2

In the scenario presented in the previous question, it transpires that only the first four data points are reliable. The correlation between the two variables will be investigated with an NHBST.

1. It is proposed to test the Null Hypothesis by carrying out a permutation test. Write down the permutations for the numbers 1 to 4.
2. With just 4 points, identify a scenario when you can you're your calculations because the Null Hypothesis has already been rejected or will not be rejected.
3. Calculate r for each permutation, noting the original value, and carry out the test by stating how many values are above the correct value.

Question 3

Let X and Y be two random variables with means μ_X, μ_Y and standard deviations σ_X, σ_Y respectively. Define the covariance of X and Y as

$$\text{cov}(X, Y) = E[\ (X - \mu_X)(Y - \mu_Y)\ ] = E[\ XY\ ] - \mu_X\mu_Y.$$

The correlation coefficient is then defined by

$$\rho_{XY} = \frac{\text{cov}(X,Y)}{\sigma_X\sigma_Y}.$$

It is required to link the variable Y to X by a linear equation $Y = \alpha + \beta X$.

1. Values of the coefficients α and β will be estimated by finding a and b respectively such that the quantity Q, defined by

   $$Q = E[(Y - a - bX)^2]$$

   is minimised with respect to a and b. Show that this leads to the equations

   $$\mu_Y = a + b\mu_X \text{ and } b = \frac{\rho_{XY}\sigma_Y}{\sigma_X}.$$

2. Show further that the value of Q at this minimum point is the value

$$E[(Y - a - bX)^2] = \sigma_Y^2(1 - \rho_{XY}^2).$$

Question 4

It is to be determined whether the fruit yield of tomato plants depends on the amount used of two additives for the soil. The regression model shown here will be applied:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2,$$

where Y is the weight of fruit yielded, X_i is the amount in appropriate units of additive i, $i = 1, 2$.

1. Write out the elements $X^{\mathrm{T}}X$ and $X^{\mathrm{T}}Y$ in the equation for the pseudo-inverse OLS estimator for (α, β_1, β_2). These elements should be written in terms of the Y_i and $X_{i,1}$, $X_{i,2}$.

Solution

Start by writing down X: let $X_{i,j}$ be the i-th value of variable j.

Similarly, Y_i is the i-th element of Y, so we have:

The matrix X is $\begin{pmatrix} 1 & X_{1,1} & X_{1,2} \\ 1 & X_{2,1} & X_{2,2} \\ \vdots & \vdots & \vdots \end{pmatrix}$.

Work out the two elements required:

Firstly $X^{\mathrm{T}}X = \begin{pmatrix} n & \sum_i X_{i,1} & \sum_i X_{i,2} \\ \sum_i X_{i,1} & \sum_i X_{i,1}^2 & \sum_i X_{i,1}X_{i,2} \\ \sum_i X_{i,2} & \sum_i X_{i,1}X_{i,2} & \sum_i X_{i,2}^2 \end{pmatrix}$ and $X^{\mathrm{T}}Y = \begin{pmatrix} \sum_i Y_i \\ \sum_i Y_i X_{i,1} \\ \sum_i Y_i X_{i,2} \end{pmatrix}$.

2. Use these results to calculate the OLS estimator b for the data given below.

Solution

Using the data given, the values of these matrices are:

$$X^{\mathrm{T}}X = \begin{pmatrix} 23 & 126.13 & 95.22 \\ 126.13 & 903.521 & 453.066 \\ 95.22 & 453.066 & 607.293 \end{pmatrix}, \; Y = \begin{pmatrix} 217.56 \\ 1594.014 \\ 593.271 \end{pmatrix}.$$

This can be solved by finding the inverse of the 3 by 3 matrix $X^{\mathrm{T}}X$ and then using this directly in the OLS equation.

Alternatively, noting that the vector of parameters b satisfies

$X^{\mathrm{T}}Xb = X^{\mathrm{T}}Y$, we can use Cramer's rule.

Either way we find that $b = (4.576, 1.59, -0.927)^{\mathrm{T}}$.

3. Set up the Null and alternative Hypothesis and other elements of the test to investigate whether the last variable is required to 'explain' Y. Carry out the test with the data given and using your previous work. Your test statistic will be

$$\frac{\|Xb\|^2 - \|Xb_s\|^2}{q} \Big/ \frac{\|e\|^2}{n-p}.$$

Solution

In conducting this test, the number of variables we are discounting, or rather whose coefficients we are setting to 0, is 1. Therefore $q = 1$ in the equation for the statistic.

The first step is to calculate the denominator expression, the value of $\|e\|^2$. This is the sum of squares of the residuals.

Therefore for each i, work out $e_i = Y_i - (4.576 + 1.59X_{i,1} - 0.927X_{i,2})$

Then sum the squares of these residual values.

The result is $\sum_i e_i^2 = 174.331$.

The denominator is $||Xb||^2 - ||Xb_s||^2$.

The first term here is the sum of the squares of the predicted values:

$||Xb||^2 = \sum_i (4.576 + 1.59X_{i,1} - 0.927X_{i,2})^2 = 2980.488$.

The next term is the same idea of a sum of squares of residuals, but now calculated from predicting Y from a simple linear model involving only variable X_1. When this is done, the parameters are $a = -0.92$ and $b = 1.893$.

Thus the estimates for i are $-0.92 + 1.893X_1$.

This sum of squares o these values is 2816.764, so

$||Xb||^2 - ||Xb_s||^2 = 163.724$.

When this is all used in the equation for the statistic we get:

$$\frac{\|Xb\|^2 - \|Xb_s\|^2}{q} \Big/ \frac{\|e\|^2}{n-p} = (163.724/1)/(174.331/20) = 18.783$$

The critical value for the statistic comes from the F distribution at 1 degree of freedom above the line and 23 – 3 below; it is 4.351. We therefore reject the Null Hypothesis.

4. Repeat this work to investigate whether both variables are required to 'explain' Y.

Solution

If we discard both variables, then the predictor for Y is simply the mean of Y. The bottom line in the expression

$$\frac{\|Xb\|^2 - \|Xb_s\|^2}{q} \Big/ \frac{\|e\|^2}{n-p}$$ is the same.

The top line is now $||Xb||^2 - \overline{Y}^2 = 2980.488 - 23\text{x}9.459^2 = 922.56$.

The same calculations as before give a test value of:

(922.56/2)/(174.331/20) = 9.392.

The critical value is 3.493 at 2 and 20 degrees of freedom, so the Null Hypothesis is rejected.

The data is given in the following table:

Y	X_1	X_2
4.75	0.58	6.98
17.42	8.39	0.71
9.29	4.67	2.47
3.29	2.78	1.99
4.89	4.14	7.57
0.33	1.15	4.71
5.99	0.95	0.09
17.7	7.44	0.31
15.87	8.68	0.53
-4.72	0.99	9.22
7.00	4.36	7.42
11.83	6.87	1.87
12.88	8.77	5.55
10.65	3.81	5.83
10.16	7.84	2.91
6.62	6.89	7.49
21.02	8.3	0.91
8.35	8.27	6.96
11.92	5.57	0.19
-1.36	0.64	6.21
19.29	9.57	2.83
3.71	6.05	9.52
20.68	9.42	2.95