MS Module 19: Correlation – practice problems

(The attached PDF file has better formatting.)

Exercise 19.1: Correlation

A regression model $Y_j = \beta_0 + \beta_1 X_j + \epsilon_j$ has N = 11 observations. The sample correlation between X and Y is 0.60. We test the null hypothesis H_0: $\rho = 0$ (the true correlation between the X and Y variables is zero).

A. What is the *t* value to test the null hypothesis?
B. What is the *p* value to test the null hypothesis?

Part A: The *t* value is $R \sqrt{(n-2)} / \sqrt{(1 - R^2)} = 0.6 \times (11 - 2)^{0.5} / (1 - 0.6^2)^{0.5} = 2.25000$

Part B: The *t* distribution has n-2 degrees of freedom, so the *p* value for a two-tailed test is 0.051 (table look-up or spread-sheet function).

Question: The sample correlation is 0.60, which is much different from zero, yet the *p* value is 5.1%, which does not satisfy even a 5% significance level.

Answer: The scenario has only 11 observations. Even if the true correlation is zero, a sample with only a few observations often shows a high sample correlation.

Exercise 19.2: Correlation and β_1

A linear regression with 11 data points has an estimated β_1 of 4.5 and a sample correlation between the X and Y values of 0.60.

A. What is the *t* value to test the null hypothesis that the correlation ρ is zero?
B. What is the *t* value to test the null hypothesis that β_1 is zero?
C. What is the standard deviation of the estimate of β_1?

Part A: The *t* value to test the null hypothesis that the correlation ρ is zero is $r\sqrt{(n-2)} / \sqrt{(1-r^2)}$ =

$$0.6 \times (11 - 2)^{0.5} / (1 - 0.6^2)^{0.5} = 2.25000$$

Part B: The *t* value to test the null hypothesis that β_1 is zero is $\hat{\beta}_1 / \sigma(\hat{\beta}_1) = 4.5 / \sigma(\hat{\beta}_1)$, where

- $\hat{\beta}_1$ is the estimate of β_1.
- $s(\hat{\beta}_1)$ is the standard deviation of the estimate of β_1.

Part C: The two tests are the same: $\rho = 0$ implies $\beta_1 = 0$ and vice versa.

$$4.5 / s(\hat{\beta}_1) = 2.25 \Rightarrow s(\hat{\beta}_1) = 4.5 / 2.25 = 2.$$

Exercise 19.3: Fisher transformation

X and Y are a bivariate normal distribution from which a sample of 40 observations is taken. The sample correlation between X and Y is 0.833. We test the null hypothesis H_0: $\rho = 0.750$. The alternative hypothesis is H_a: $\rho_o > 0.750$.

A. What is the Fisher transform V of the random variable R of the correlation R between X and Y?
B. What is the Fisher transform of the observed correlation ρ?
C. What is the distribution of V?
D. What is the Fisher transform of the correlation ρ_0 assumed in the null hypothesis?
E. What is the *z* value to test this null hypothesis?
F. What is the *p* value for this test of the null hypothesis?
G. What is the 95% confidence interval for the true value of the Fisher transform of the correlation?
H. What is the 95% confidence interval for the true value of the correlation?

Part A: The Fisher transform V = ½ *ln*((1+R)/(1-R))

Question: Why do we use a Fisher transform of the sample correlation?

Answer: The correlation is bounded by –1 and +1, and it is not normally distributed. If the sample correlation is from a bivariate normal distribution, the Fisher transform of the correlation has (approximately) a normal distribution with a mean of ½ *ln*((1+R)/(1-R)) and a variance of 1/(n-3). The textbook notes that "the rationale for the transformation is to obtain a function of R that has a variance independent of r; this would not be the case with R itself."

Part B: The Fisher transform of $\rho = 0.833$ is $0.5 \times \ln((1 + 0.833) / (1 – 0.833)) = 1.197858$

Part C: V has a normal distribution with mean $\mu = 1.19786$ and variance $\sigma^2 = 1/(40-3) = 1/37$.

Part D: The Fisher transform of $\rho = 0.75$ is $0.5 \times \ln((1 + 0.75) / (1 – 0.75)) = 0.972955$

Part E: The *z* value is $(\rho – \rho_o)$ / standard deviation of V = $(1.19786 – 0.97296) / (1 / 37^{0.5}) = 1.36801$

Part F: The *p* value for a one-sided test is 1 – CDF(1.368) = 0.08565. The *p* value for a two sided test is twice this value.

Part G: The *z* value for a two-sided 95% confidence interval is 1.96. The 95% confidence interval is

$(1.19786 – 1.96 / 37^{0.5}, 1.19786 + 1.96 / 37^{0.5}) = (0.87564, 1.52008)$.

Part H: The inverse of the Fisher transform is $(e^{2x} – 1) / (e^{2x} + 1)$.

- $(\exp(2 \times 0.87564) – 1) / (\exp(2 \times 0.87564) + 1) = 0.70423$
- $(\exp(2 \times 1.52008) – 1) / (\exp(2 \times 1.52008) + 1) = 0.90871$

The 95% confidence interval for the correlation is (0.70423, 0.90871).

Question: The observed correlation 0.833 is not the midpoint of this confidence interval.

Answer: The correlation does not have a normal distribution. Suppose we had a sample of four points with a sample correlation of 0.980. With only four points, the observed correlation may be due to chance. The 95% confidence interval is wide, but the correlation can not be greater than 1.000. The confidence interval of the Fisher transform is symmetric, and the Fisher transform of the observed correlation is the mid-point of the confidence interval.

Question: Do exam problems give the sample correlation or must we derive it?

Answer: Exam problems give either the sample correlation or data from which it derive it. The data may be

- Summary statistics (sums, sums of squares, and sum of cross products)
- S_{xx}, S_{yy}, and S_{xy}
- The total sum of squares (SST) and the error sum of squares (SSE).