Analysis of the Correlation Coefficient between Math Placement Exam scores and Grades (Pass/Fail)

1 Introduction

Suppose we have a sample of $N$ students, who take the Math Placement Exam. Let $x_i$ be the score of the $i$th student. Further, let the grade $y_i$ of the $i$th student in Calculus be either a 1 (pass) or 0 (fail).

Let the mean score of the sample be denoted by $\mu$. Let the mean score of the students who fail be denoted by $\mu_F$ and those who pass, by $\mu_P$. Let the standard deviation of the scores (for the entire sample) be denoted by $\sigma_x$, and the standard deviation of the grades be denoted by $\sigma_y$. The mean of the grades $y_i$ is given by $N_P/N$, which is the ratio of students who pass.

Our goal is to estimate the correlation coefficient between the MPE scores $x_i$ and the grades $y_i$.

The correlation coefficient between the test scores $x_i$ and the grades $y_i$ is defined by the formula

$$R = \sum_{i=1}^{N} \frac{(x_i - \bar{x})(y_i - \bar{y})}{(N-1)\sigma_x\sigma_y} \quad (1)$$

We can split the population into two components: those who pass ($x_i^P, y_i^P = 1$) and those who fail ($x_i^F, y_i^F = 0$). Note that the mean of the grades (pass/fail) is given by $\mu_y = N_P/N = \bar{y}$ which implies $N_F/N = (1 - \bar{y})$.

$$R = \sum_{i=1}^{N_P} \frac{(x_i^P - \mu)(1 - \bar{y})}{(N-1)\sigma_x\sigma_y} + \sum_{i=1}^{N_F} \frac{(x_i^F - \mu)(0 - \bar{y})}{(N-1)\sigma_x\sigma_y}$$

$$= \frac{N_P}{N-1} \sum_{i=1}^{N_P} \frac{(x_i^P - \mu)}{N_P\sigma_x\sigma_y} + \frac{N_F}{N-1} \sum_{i=1}^{N_F} \frac{(x_i^F - \mu)}{N_F\sigma_x\sigma_y}$$

$$= \frac{N_P}{N-1} \frac{(1 - \bar{y})\mu_P - \mu}{\sigma_x\sigma_y} + \frac{N_F}{N-1} \frac{(0 - \bar{y})\mu_F - \mu}{\sigma_x\sigma_y}$$

$$= \frac{N}{N-1} \left( \bar{y}(1 - \bar{y})\frac{\mu_P - \mu}{\sigma_x\sigma_y} + (1 - \bar{y})(0 - \bar{y})\frac{\mu_F - \mu}{\sigma_x\sigma_y} \right)$$

$$= \frac{N}{N-1} \bar{y}(1 - \bar{y}) \left( \frac{\mu_P - \mu}{\sigma_x\sigma_y} - \frac{\mu_F - \mu}{\sigma_x\sigma_y} \right)$$

$$= \frac{N}{N-1} \bar{y}(1 - \bar{y}) \Delta$$

where $\Delta = \mu_P - \mu_F$ is the difference between the passing and failing students’ mean scores.
To estimate this further, we calculate $\sigma_y$ by

$$\sigma_y^2 = \sum_{i=1}^{N} \frac{(y_i - \bar{y})^2}{N - 1}$$

$$= \sum_{i=1}^{N_P} \frac{(1 - \bar{y})^2}{N - 1} + \sum_{i=1}^{N_F} \frac{(0 - \bar{y})^2}{N - 1}$$

$$= \frac{N}{N - 1} \left( \sum_{i=1}^{N_P} \frac{(1 - \bar{y})^2}{N} + \sum_{i=1}^{N_F} \frac{(0 - \bar{y})^2}{N} \right)$$

$$= \frac{N}{N - 1} \left( \frac{N_P}{N} (1 - \bar{y})^2 + \frac{N_F}{N} (0 - \bar{y})^2 \right)$$

$$= \frac{N}{N - 1} \left( (1 - \bar{y})^2 \bar{y} + \bar{y}^2 (1 - \bar{y}) \right)$$

$$= \bar{y} (1 - \bar{y}) \frac{N}{N - 1}$$

The correlation coefficient simplifies to

$$R = \frac{\sigma_y^2}{\sigma_x \sigma_y} (\mu_P - \mu_F)$$

$$= \frac{\sigma_y}{\sigma_x} (\mu_P - \mu_F)$$

$$= \frac{\sqrt{\bar{y} (1 - \bar{y})}}{\sigma_x} \frac{\sqrt{N}}{\sqrt{N - 1}} \Delta$$

Asymptotically, as $N \to \infty$,

$$R \to \frac{\sqrt{\bar{y} (1 - \bar{y})}}{\sigma_x} \Delta$$

In the 2007 data set, $\sigma_x = 6.336, \bar{y} = 0.69988, \Delta = 5.7725$ which gives $R \approx 0.41755$ (neglecting the $\sqrt{N/(N - 1)}$ factor. The actual correlation coefficient is 0.417673. Regardless of the distribution of pass/fail grades, the regression coefficient is bounded above by $\Delta/(2\sigma_x)$. The values for $\Delta/\sigma_x$ are almost constant, regardless of which questions are chosen.
This implies that the correlation coefficient is not a good way to analyze discrete ordinal data.

We look at two limiting cases. In the first case, consider the situation when the MPE scores are distributed normally, with $\mu = 0$ and $\sigma = 1$. Assume everyone who scores above the mean passes, and everyone below the mean fails. We compute $\mu_P$ by

$$\mu_P = \frac{1}{\sqrt{2\pi}} \int_0^\infty xe^{-x^2/2}dx = \frac{1}{\sqrt{2\pi}}$$

Since $\mu_F = -\mu_P$ we have $\Delta = \frac{2}{\sqrt{2\pi}}$ and

$$R = \sqrt{(1/2)(1/2)} \frac{2}{\sqrt{2\pi}} = \frac{\sqrt{2}}{\sqrt{\pi}} = 0.7979$$

In the second case, we assume we have uniformly distributed scores on the interval $[0, L]$, and anyone scoring less than $L/2$ fails, and anyone above $L/2$ passes. For this case $\mu = L/2$, $\mu_P = 3L/4$, $\mu_F = L/4$ and $\sigma = L/\sqrt{12}$. This leads to

$$R = \sqrt{(1/2)(1/2)} \frac{L/2}{L/\sqrt{12}} = \sqrt{3/4} = 0.866$$

In both of these cases we have **PERFECT** prediction – anyone below the mean fails, anyone above passes. But, the $R$ values range from a low of 0.4 in the case of normally distributed scores to a high of 0.866 in the case of uniformly distributed scores.

When we say the $R^2$ explains a certain percentage of the variability, we need to say that it is with respect to a linear regression model.