14.8: Multiple Regression

Multiple regression is to the linear regression we just covered as one-way ANOVA is to m-way ANOVA. In m-way ANOVA we have one DV and m discrete IVs. With multiple regression we have one DV (univariate) and k continuous IVs. We will label the DV with y and the IVs with $x_1, x_2, \ldots, x_k$. The idea is to predict y with y' via

$$y' = a + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k \tag{14.8.1}$$

or, using summation notation

$$y' = a + \sum_{j=1}^{k} b_j x_j \tag{14.8.2}$$

Sometimes we (and SPSS) write $a = b_0$. The explicit formula for the coefficients a and b_j are long so we won't give them here but, instead, we will rely on SPSS to compute the coefficients for us. Just the same, we should remember that the coefficients are computed using the least squares method, where the sum of the squared deviations is minimized. That is, a and the b_j are such that

$$\begin{aligned} E &= \sum_{i=1}^{n} (y_i - y_i')^2 \\ &= \sum_{i=1}^{n} (y_i - [a + \sum_{j=1}^{k} b_j x_{ji}])^2 \end{aligned}$$

is minimized. (Here we are using $(y_i, x_{1i}, x_{2i}, \ldots, x_{ki})$ to represent data point i.) If you like calculus and have a few minutes to spare, the equations for a and the b_j can be found by solving:

$$\frac{\partial E}{\partial a} = 0, \quad \frac{\partial E}{\partial b_1} = 0, \quad \cdots \quad \frac{\partial E}{\partial b_k} = 0 \tag{14.8.3}$$

for a and the b_j. The result will contain al the familiar terms like $\sum y$, $\sum y x_j$, etc. It also turns out that the "normal equations" for a and the b_j that result have a pattern that can be captured with a simple linear algebra equation that we will see in Chapter 17.

Some terminology: the b_j (including b_0) are known as *partial regression coefficients*.

14.10.1: Multiple regression coefficient, *r*

An overall correlation coefficient, r, can be computed using pairwise bivariate correlation coefficients as defined in the previous Section 14.2. This overall correlation is defined as $r = r_{y'y}$, the bivariate correlation coefficient of the predicted values y' versus the data y. For the case of 2 IVs, the formula is

$$r = \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 - 2 r_{yx_1} r_{yx_2} r_{x_1 x_2}}{1 - r_{x_1 x_2}^2}} \tag{14.8.4}$$

where r_{yx_1} is the bivariate correlation coefficient between y and x_1, etc. It is true that $-1 \leq r \leq 1$ as with the bivariate r.

Example 14.6 : Suppose that you have used SPSS to obtain the regression equation

$$y' = -44.572 + 87.679 x_1 + 14.519 x_2 \tag{14.8.5}$$

for the following data :

Student	GPA, x_1	Age, x_2	Score, y	x_1^2	x_2^2	y^2	$x_1 y$	$x_2 y$	$x_1 x_2$
A	3.2	22	550	10.24	484	302500	1760	12100	70.4
B	2.7	27	570	7.29	729	324900	1539	15390	72.9
C	2.5	24	525	6.25	576	275625	1312.5	12600	60
D	3.4	28	670	11.56	784	448900	2278	18760	95.2

E	2.2	23	490	4.84	529	240100	1078	11270	50.6
$n = 5$	$\sum x_1 =$ 14	$\sum x_2 =$ 124	$\sum y =$ 2805	$\sum x_1^2 =$ 40.18	$\sum x_2^2 =$ 3102	$\sum y^2 =$ 1592025	$\sum x_1 y =$ 7967.5	$\sum x_2 y =$ 70120	$\sum x_1 x_2 =$ 349.1

Compute the multiple correlation coefficient.

Solution :

First we need to compute the pairwise correlations $r_{x_1 y}$, $r_{x_2 y}$, and $r_{x_1 x_2}$. (Note that $r_{x_1 y}$ = $r_{y x_1}$, etc. because the correlation matrix is symmetric.)

$$\begin{aligned} r_{x_1 y} &= \frac{n(\sum x_1 y) - (\sum x_1)(\sum y)}{\sqrt{[n(\sum x_1^2) - (\sum x_1)^2][n(\sum y^2) - (\sum y)^2]}} \\ &= \frac{5(7967.5) - (14)(2805)}{\sqrt{[5(40.18) - (14)^2][5(1592025) - (2805)^2]}} \\ &= 0.845 \end{aligned}$$

$$\begin{aligned} r_{x_2 y} &= \frac{n(\sum x_2 y) - (\sum x_2)(\sum y)}{\sqrt{[n(\sum x_2^2) - (\sum x_2)^2][n(\sum y^2) - (\sum y)^2]}} \\ &= \frac{5(70120) - (124)(2805)}{\sqrt{[5(3102) - (124)^2][5(1592025) - (2805)^2]}} \\ &= 0.791 \end{aligned}$$

$$\begin{aligned} r_{x_1 x_2} &= \frac{n(\sum x_1 x_2) - (\sum x_1)(\sum x_2)}{\sqrt{[n(\sum x_1^2) - (\sum x_1)^2][n(\sum x_2^2) - (\sum x_2)^2]}} \\ &= \frac{5(349.1) - (14)(124)}{\sqrt{[5(40.18) - (14)^2][5(3102) - (124)^2]}} \\ &= 0.371 \end{aligned}$$

Now use these in :

$$\begin{aligned} r &= \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 - 2 r_{yx_1} r_{yx_2} r_{x_1 x_2}}{1 - r_{x_1 x_2}^2}} \\ &= \sqrt{\frac{(0.845)^2 + (0.791)^2 - (2)(0.845)(0.791)(0.371)}{1 - (0.371)^2}} \\ &= 0.989 \end{aligned}$$

14.10.2: Significance of *r*

Here we want to test the hypotheses :

$$\begin{aligned} H_0 &: \rho = 0 \\ H_1 &: \rho \neq 0 \end{aligned}$$

where ρ is the population multiple regression correlation coefficient.

To test the hypothesis we use

Rendered by QuickLaTeX.com

with

$$\nu_1 = n-k \backslash \ \backslash \ \backslash \ (\text{d.f.N.}) \backslash \ \backslash \ \backslash \ \backslash \ \text{and} \backslash \ \backslash \ \backslash \ \backslash \ \nu_2 = n-k-1 \backslash \ \backslash \ \backslash \ (\text{d.f.D.}) \tag{14.8.6}$$

here:

$$\begin{aligned} n = &\ \text{sample size} \\ k = &\ \text{number of IVs} \\ r = &\ \text{multiple correlation coefficient} \end{aligned}$$

(Note: This “F-test” is similar to but not the same as the “ANOVA” output given by SPSS when you run a regression.)

Example 14.7 : Continuing with Example 14.6, test the significance of r.

Solution :

1. Hypotheses.

$$\begin{aligned} H_0 &: \rho = 0 \\ H_1 &: \rho \neq 0 \end{aligned}$$

2. Critical statistic. From the **Rank Correlation Coefficient Critical Values Table** (i.e., the critical values for the Spearman correlation) with

$$\begin{aligned} \nu_1 = &\ n-k = 5-2 = 3 \\ \nu_2 = &\ n-k-1 = 5-2-1 = 2 \\ \alpha = &\ 0.05 \end{aligned}$$

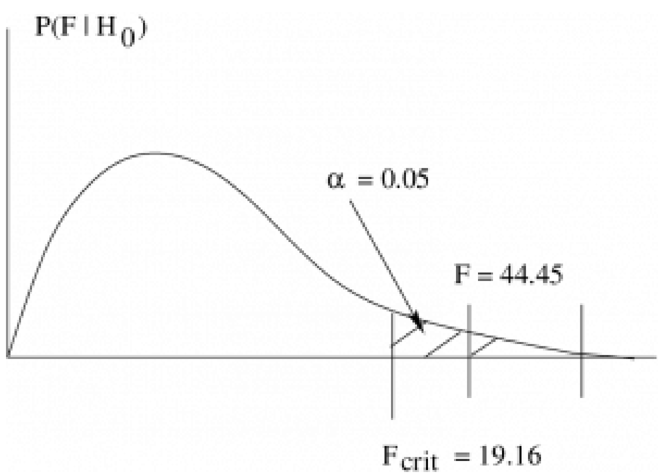
find

$$F_{\text{crit}} = 19.16 \tag{14.8.7}$$

3. Test statistic.

Rendered by QuickLaTeX.com

4. Decision.



Reject H_0.

5. Interpretation.

$r = 0.989$ is significant.

14.10.3: Other descriptions of correlation

1. Coefficient of multiple determination: r^2. This quantity still has the interpretation as fraction of variance explained by the (multiple regression) model.

2. Adjusted r^2:

$$r^2_{\text{adj}} = 1 - \left[\frac{(1-r^2)(n-1)}{n-k-1}\right] \tag{14.8.8}$$

r^2_{adj} gives a better (unbiased) estimate of the population value for ρ^2 by correcting for degrees of freedom just as the sample s^2 with its degrees of freedom equal to $n-1$ gives an unbiased estimate of the population σ^2.

Example 14.8 : Continuing Example 14.6, we had $r = 0.989$ so

$$r^2 = 0.978 \tag{14.8.9}$$

and

$$\begin{aligned} r^2_{\text{adj}} &= 1 - \left[\frac{(1-r^2)(n-1)}{n-k-1}\right] \\ r^2_{\text{adj}} &= 1 - \left[\frac{(1-0.978)(5-1)}{5-2-1}\right] \\ &= 0.956 \end{aligned}$$

▢
