Multiple Correlation

As seen earlier, the proportion of variation explained by the linear regression of Y on the regressors $X_1, \ldots, X_{p-1}$ is given by

$$R^2 = \frac{\text{SS}_{reg}}{\text{SST (corrected)}} = 1 - \frac{\text{RSS}}{\text{SST (corrected)}} = 1 - \frac{Y'(I-P)Y}{Y'(I - \frac{1}{n}\mathbf{1}\mathbf{1}')Y}.$$

Consider simple linear regression: Then $p = 2$ and $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

$$\text{RSS} = \sum_{i=1}^n (y_i - \bar{y})^2 - \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2,$$

so that

$$\text{SS}_{reg} = \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{\left\{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})\right\}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Therefore,

$$\begin{aligned} R^2 &= \frac{\text{SS}_{reg}}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\left\{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})\right\}^2}{\left\{\sum_{i=1}^n (x_i - \bar{x})^2\right\}\left\{\sum_{i=1}^n (y_i - \bar{y})^2\right\}} \\ &= \left\{ \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left\{\sum_{i=1}^n (x_i - \bar{x})^2\right\}\left\{\sum_{i=1}^n (y_i - \bar{y})^2\right\}}} \right\}^2 = r_{XY}^2, \end{aligned}$$

where

$$\begin{aligned} r_{XY} &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \\ &= \text{sample correlation coefficient between } X \text{ and } Y. \end{aligned}$$

This connection between R^2 and r^2 is intuitively meaningful since a good linear fit is related to a good linear association between X and Y. What happens when there are multiple regressors, $X_1, X_2, \ldots, X_{p-1}$?

We define the *multiple correlation coefficient* between Y and $X_1, \ldots, X_{p-1}$ as the *maximum* correlation coefficient between Y and any linear function of $X_1, \ldots, X_{p-1} = \max_{\mathbf{a}} Corr(Y, a_0 + a_1 X_1 + \cdots + a_{p-1} X_{p-1}) = R^*$ (say).

If $Cov\left(\left(\begin{array}{c} Y \\ \mathbf{X} \end{array}\right)\right) = \left(\begin{array}{cc} \sigma_{YY} & \sigma'_{XY} \\ \sigma_{XY} & \Sigma_X \end{array}\right)$, then

$$Corr^2(Y, a'X) = \frac{Cov^2(Y, a'X)}{Var(Y)Var(a'X)} = \frac{\{a'Cov(Y, X)\}^2}{Var(Y)Var(a'X)} = \frac{\{a'\sigma_{XY}\}^2}{\sigma_{YY}a'\Sigma_X a}.$$

Further, taking $u' = a'\Sigma_X^{1/2}$ and $v = \Sigma_X^{-1/2}\sigma_{XY}$,

$$\begin{aligned} \frac{a'\sigma_{XY}}{(\sigma_{YY}a'\Sigma_X a)^{1/2}} &= \frac{a'\Sigma_X^{1/2}\Sigma_X^{-1/2}\sigma_{XY}}{(\sigma_{YY}a'\Sigma_X a)^{1/2}} = \frac{u'v}{(\sigma_{YY}a'\Sigma_X a)^{1/2}} \\ &\le \frac{(u'u)^{1/2}(v'v)^{1/2}}{(\sigma_{YY}a'\Sigma_X a)^{1/2}} = \frac{(a'\Sigma_X a)^{1/2}\left(\sigma'_{XY}\Sigma_X^{-1}\sigma_{XY}\right)^{1/2}}{(\sigma_{YY}a'\Sigma_X a)^{1/2}} \\ &= \left(\frac{\sigma'_{XY}\Sigma_X^{-1}\sigma_{XY}}{\sigma_{YY}}\right)^{1/2}, \end{aligned}$$

with equality if we take $u \propto v$ or $a = \Sigma_X^{-1}\sigma_{XY}$. Since $R^* = \sqrt{\sigma'_{XY}\Sigma_X^{-1}\sigma_{XY}/\sigma_{YY}}$, $0 \le R^* \le 1$ unlike the ordinary correlation coefficient. Now let us see why $(R^*)^2$ (square of multiple correlation coefficient) is the same as the coefficient of determination, R^2 (proportion of variability explained by the regressors). Suppose

$$\left(\begin{array}{c} Y \\ \mathbf{X} \end{array}\right) \sim N\left(\left(\begin{array}{c} \mu_Y \\ \mu_X \end{array}\right), \left(\begin{array}{cc} \sigma_{YY} & \sigma'_{XY} \\ \sigma_{XY} & \Sigma_X \end{array}\right)\right).$$

Then,

$$Y|\mathbf{X} \sim N\left(\mu_Y + \sigma'_{XY}\Sigma_X^{-1}(\mathbf{X} - \mu_X), \sigma_{YY} - \sigma'_{XY}\Sigma_X^{-1}\sigma_{XY}\right).$$

Thus, $E(Y|\mathbf{X}) = \mu_Y - \sigma'_{XY}\Sigma_X^{-1}\mu_X + \sigma'_{XY}\Sigma_X^{-1}\mathbf{X}$ and
$Var(Y|\mathbf{X}) = \sigma_{YY} - \sigma'_{XY}\Sigma_X^{-1}\sigma_{XY})$. Therefore,

$$\begin{aligned} Corr(Y, E(Y|\mathbf{X})) &= \frac{Cov(Y, \sigma'_{XY}\Sigma_X^{-1}\mathbf{X})}{\sqrt{\sigma_{YY}\sigma'_{XY}\Sigma_X^{-1}\Sigma_X\Sigma_X^{-1}\sigma_{XY}}} \\ &= \frac{\sigma'_{XY}\Sigma_X^{-1}\sigma_{XY}}{\sqrt{\sigma_{YY}}\sqrt{\sigma'_{XY}\Sigma_X^{-1}\sigma_{XY}}} = R^*. \end{aligned}$$

i.e., R^* = correlation coefficient between Y and the conditional expectation of $Y|\mathbf{X}$ (or the regression of Y on $\mathbf{X}$, when the conditional expectation is linear). Further, $Var(Y) - E\left(Var(Y|\mathbf{X})\right) = \sigma_{YY} - \left(\sigma_{YY} - \sigma'_{XY}\Sigma_X^{-1}\sigma_{XY}\right) = \sigma'_{XY}\Sigma_X^{-1}\sigma_{XY}$, so that the proportion of variation in Y explained by the regression on $\mathbf{X}$ is equal to

$$R^2 = \frac{Var(Y) - E\left(Var(Y|\mathbf{X})\right)}{Var(Y)} = \frac{\sigma'_{XY}\Sigma_X^{-1}\sigma_{XY}}{\sigma_{YY}} = (R^*)^2.$$