Adjusted R^2

1.1 Setting

Consider the linear model:

$$Y = X\beta + \epsilon, \tag{1.1.1}$$

where Y is an $n \times 1$ outcome, X is an $n \times k$ design matrix, assumed to include an intercept, and $\epsilon \sim N(0, \sigma^2 I)$ is an $n \times 1$ residual vector. Model (1.1.1) is described as the *full-model*, in contrast to the *reduced-model*, which includes an intercept only:

$$Y = 1\beta_0 + \varepsilon \tag{1.1.2}$$

1.2 Sum of Squares Decomposition

The projection matrix for the full model is $P_X = X(X'X)^{-1}X'$, and that for the reduced model is $P_0 = 1(1'1)^{-1}1'$. The projection of Y onto X is $\hat{Y}_X = P_X Y$, and that onto 1 is $\hat{Y}_0 = P_0 Y$. The **total sum of squares** is defined as:

$$||Y - \hat{Y}_0||^2 = ||(I - P_0)Y||^2 = Y'(I - P_0)Y.$$

Since $Y - \hat{Y} \in \text{im}(X)^{\perp}$ and $\hat{Y} - \hat{Y}_0 \in \text{im}(X)$, the total sum of squares decomposes as:

$$\begin{aligned} ||Y - \hat{Y}_0||^2 &= ||(I - P_0)Y||^2 \\ &= ||(I - P_X + P_X - P_0)Y||^2 \\ &= ||(I - P_X)Y||^2 + ||(P_X - P_0)Y||^2 \\ &= ||Y - \hat{Y}_X||^2 + ||\hat{Y}_X - \hat{Y}_0||^2. \end{aligned}$$

Here $||Y - \hat{Y}_X||^2 = Y'(I - P_X)Y$ is the **residual sum of squares** while $||\hat{Y}_X - \hat{Y}_0||^2 = Y'(P_X - P_0)Y$ is the **model sum of squares**.

1.3 Coefficient of Determination

The **coefficient of determination** for the full model (1.1.1) is defined as:

$$R^2 = \frac{||\hat{Y}_X - \hat{Y}_0||^2}{||Y - \hat{Y}_0||^2}.$$

This is the proportion of total variation explained by the columns of X other than the intercept. Note that:

$$1 - R^2 = \frac{||Y - \hat{Y}_X||^2}{||Y - \hat{Y}_0||^2}$$

1.4 Snedecor's Statistic

The ***F*-statistic** comparing the full (1.1.1) and reduced (1.1.2) models is:

$$T_F = \frac{||\hat{Y}_X - \hat{Y}_0||^2/(k-1)}{||Y - \hat{Y}_X||/(n-k)} \overset{H_0}{\sim} F_{k-1,n-k}(0).$$

Under the null hypothesis $\mathbb{E}(Y) \in \text{im}(1)$, T_F follows a central F distribution with numerator and denominator degrees of freedom $k-1$ and $n-k$.

1.5 Distribution of R^2

The F-statistic may be expressed in terms of the coefficient of determination:

$$T_F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}.$$

Likewise, R^2 may be expressed using the F-statistic:

$$R^2 = \frac{(k-1)T_F}{(k-1)T_F + (n-k)}.$$

For $T_F \sim F_{\nu_1,n_2}(0)$, $\nu_1 = k-1$, $\nu_2 = n-k$, the random variable $\nu_1 T_F/(\nu_1 T_F + \nu_2)$ follows a beta distribution with parameters $\alpha = \nu_1/2$ and $\beta = \nu_2/2$.

1.6 Adjusted R^2

Now, under H_0, $R^2 \sim B(\nu_1/2, \nu_2/2)$, and has expectation:

$$\mathbb{E}(R^2) = \frac{\nu_1}{\nu_1 + \nu_2} = \frac{k-1}{n-1}.$$

However, the expected value of R^2 is zero. Thus, R^2 is upward biased in general. To correct for this, consider the **adjusted $\boldsymbol{R^2}$**, defined as:

$$R_a^2 = R^2 + (1-R^2)\frac{(k-1)}{(n-k)}.$$

Observe that, in contrast to R^2, R_a^2 is unbiased for zero under H_0:

$$\begin{aligned}
\mathbb{E}(R_a^2) &= \frac{k-1}{n-1} + \left(1 - \frac{k-1}{n-1}\right)\frac{(k-1)}{(n-k)} \\
&= \frac{k-1}{n-1} + \left\{\frac{(n-1)-(k-1)}{n-1}\right\}\frac{(k-1)}{(n-k)} = 0.
\end{aligned}$$