STOR 664 CWE 2018

(1) Variance inflation factor (**VIF**) is a measure for multi-collinearity that can be defined, in what follows, by **D1** or **D2**.

Let $A = \begin{pmatrix} a & B \\ B^t & C \end{pmatrix}$ be a block matrix in which $a > 0$, C is a $m \times m$ symmetric matrix, the $(m+1) \times (m+1)$ inverse matrix A^{-1} can be expressed as a block matrix with the (upper-left) entry $A_{11}^{-1} = (a - BC^{-1}B^t)^{-1}$. For the regression model $y = X\beta + \epsilon$, we focus on the VIF corresponding to the predictor $x_j = (x_{1j} \cdots x_{nj})^t$ in the design matrix X. Without loss of generality, let us write $X = (x_j \ X_{-j})$ where X_{-j} is the $n \times (p-1)$ submatrix of X without the column vector x_j. Therefore, $X^tX = \begin{pmatrix} x_j^t x_j & x_j^t X_{-j} \\ & \\ X_{-j}^t x_j & X_{-j}^t X_{-j} \end{pmatrix}$ which enjoys a similar form to the matrix A given above. Hence the upper-left entry of $(X^tX)^{-1}$ is defined as the j-th VIF by

$$(\mathbf{D1}) \quad V_j = [x_j^t x_j - x_j^t X_{-j} \ (X_{-j}^t X_{-j})^{-1} X_{-j}^t x_j]^{-1}$$

where the covariate vectors x_j in X are standardized to satisfy $\overline{x}_j = 0$ (zero average) and $\|x_j\|^2 = 1$ (unit length).

An alternative definition for the j-th VIF is given by

$$(\mathbf{D2}) \quad V_j = (1 - R_j^2)^{-1}$$

where R_j^2 is the coefficient of determination $R^2 = \frac{SSR}{SSTO}$ when regressing x_j on those column vectors in X_{-j}.

Question: Are (**D1**) and (**D2**) equivalent? If so, prove it. If not, how they differ?

(2) Consider the regression model $y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \ i = 1, ..., n$ where $\epsilon_i, \ i = 1, ..., n$ are iid random variables with mean 0 and variance σ^2. Assume all covariates x_{ij} are known constants, and $\beta_j, \ j = 0, 1, 2$ and σ^2 are unknown parameters.

(2a) Suppose $\epsilon_1, ..., \epsilon_n$ are iid $N(0, \sigma^2)$ random variables. Consider two reduced models M1 and M2 under the constraints $\beta_1 = 0$ and $\beta_2 = 0$ respectively. How would you choose between M1 and M2? Describe a detailed model selection procedure based on the observations $y_1, ..., y_n$.

(2b) Assume a modest sample size $n \approx 30$, say. Suppose errors $\epsilon_1, ..., \epsilon_n$ are not from a normal population and it is not clear what distribution the population follows. How would you conduct a test for the hypothesis $H_0 : \beta_1 = 0$? Provide details in your proposal.

(2c) Under the same condition in (2b) but without any constraints on β_1 and β_2, how would you construct a 95% CI for β_2?