Linear Least Squares Curve Fitting (Regression) Simulations -II

The Excel file **data_records_2021** contains 5 sets of y measurements (***y*1, *y*2,..., *y*5**) obtained for $-10 \leq x \leq +10$.

Use 5-fold cross validation in conjunction with the residual sum of squares (RSS) (see eqn 1 below) and R2 (coefficient of determination, eqn 2 below) metrics to estimate a polynomial model that best explains the given data. Show plots and statistics as appropriate to justify your selection. Also show the coefficients of the chosen polynomial model.

RSS: $\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$ (1)

R2: $1 - \frac{RSS}{\sum_{i=1}^{m}(y_i - \hat{\mu})^2} = 1 - \frac{\overline{RSS}}{\widehat{Var}(y)}$ (2)

In Eqn (2) $\hat{\mu}$ is the estimated mean and $\widehat{Var}(y)$ is the estimated variance of a data record, $\{y_i, i = 1, 2, \ldots, m\}$.

(Note: RSS and R2 can be used to evaluate performance on both, training data and test data)

Be alert to possible numerical errors arising out of ill-conditioning of $\boldsymbol{A}^T\boldsymbol{A}$ which may affect solution accuracy of the normal equations, especially for larger model orders. As before, use only the basic tools available for matrix multiplication, transpose, inverse, solution of equations, etc. (this way you may learn a lot) and not use any built-in functions that your programming environment may offer to directly solve the problem (this way you may not learn a lot).