Thema\_3

1. The provided code begins by generating an empty data frame called my data with 7 columns all (y, x1….x6) all numeric, it then precedes to generate 4 rows, using a loop. During each iteration, generates a row of random values between 1 and 10 using the “runif” function. Each row corresponds to a different observation. (in the code: 7 being the number of random values to generate 1 and 10 being the minimum and maximum values possible). It then fits a linear model where “y”, the first column, is the dependent variable and all other columns are used as independent variables (in the code: this is indicated using the dot “.”). Finally, the codes print’s the coefficients of the linear model.
2. The value of the coefficient associated with a variable gives us an indication of how that independent variable is related to the dependent variable. The direction (positive or negative) of the coefficient indicates the direction of the relationship between the independent and dependent variables. The magnitude of the coefficients gives us an idea of the importance of each variable in the prediction. It is important to note that the scale of the independent variables can affect the scale of the coefficients in a linear regression model. When the scale of the variable’s changes, the scale of the coefficients can change proportionally. This is because the coefficients represent the change in the dependent variable for a one-unit change in the corresponding independent variable. In our case all independent variables as well as the dependent variable take values in a range of 1-10 hence the coefficients, will always be within the 1-10 range, consistent with the scale of your variables. Finally, we notice that no matter how many times we run the model the coefficients for x4, x5 and x6 always come back as NA. The consistent occurrence of NA values for x4, x5, and x6 in the linear regression model is attributed to perfect multicollinearity within the dataset. This multicollinearity leads to a singular or rank-deficient model matrix, rendering the coefficients for these variables undefined. The correlation matrix highlights a notable correlation among x4, x5, and x6, and the condition number of the design matrix indicates near singularity. This effect could persist, even if the numbers are generated at random each time we run the code, due to the small sample size which leads to unstable estimations, we could also explore regularization techniques such as ridge regression to mitigate the impact of multicollinearity or examine the data for patterns in the values of x4, x5 and x6.
3. Based on the results of the executed program and their interpretation, we can conclude that the method of least squares is sensitive when perfect multicollinearity exists among variables. Multicollinearity leads to an almost singular estimation matrix, resulting in the inability to determine certain coefficients. This could lead to issues such as the inability to estimate specific coefficients and an increased risk of overfitting.

Mathematically, the matrix , represents the result of multiplying the transpose of the matrix , which contains the independent variables of a linear regression model, by , itself. In OLS this is done to compute the variance-covariance matrix of coefficient estimates, detecting multicollinearity through the condition number and most importantly to obtain estimated coefficients through . Here X is the matrix of independent variables, y is the vector of dependent variable values, is the transpose of, and is the inverse of the matrix product of . Overall, the above expression arises from solving the normal equations in the method of least squares. The objective is to minimize the sum of squared residuals, representing the squared differences between observed and predicted values. To find the coefficients minimizing this sum, the calculus of variations is used. By setting the derivative of the sum of squared residuals with respect to each coefficient to zero, a system of linear equations known as the normal equations is obtained. Solving these equations involves the matrix , and its inverse . The multiplication of by the vector y yields the coefficients that minimize the sum of squared differences, providing an optimal linear model. Hence it becomes clear that in the presence of multicollinearity , may not be invertible, this renders the calculation of the coefficient estimates impossible and without a unique estimation of coefficients, the solution is not uniquely determined.