**Block 4: Regression**

Analyse how gene expression levels in specific metabolic pathways are related to each other. To do so, make use of linear regression models where you regress gene expression levels with each other. Make a model using one gene as the dependent variable while using the other genes as predictors/independent variables. Generate regression models for a given gene using expression data from both tissues (normal and tumour) for the breast cancer part of the data.

(A) The Krebs cycle contains the following genes: EHHADH, PCCB, OXCT1, MCCC1, GCDH. Create multiple linear regression models that predict genes from the Krebs cycle using the other genes from this cycle (a total of 5 models).

Using these models, answer at least the following questions:

1. Give R2 values for these models.
2. What variables are significant in the models?
3. How would you rate the predictive properties of these models?
4. Is there a collinearity problem in any of these models?
5. Consider whether it would make sense to log-transform the data before performing the regression. How do your results change when you apply this transformation?

(B) Using lasso regression, create models that predict each gene of the Krebs cycle using the information of all genes. (Make sure the gene you are predicting is not in the set of predictors!). Look, at least, at the following aspects:

1. Use cross validation and calculate Q2values to assess predictive properties of your lasso models.
2. Discuss the results from the models. What genes are most predictive?

Create a comprehensible report and include where possible, nice figures.