SS2.1 – Regression Analysis
Lab Session

(a) Simple linear regression

The variables in the milk production data are:

current: current month milk production in pounds
previous: previous month milk production in pounds
fat: percent of fat in milk
protein: percent of protein in milk
days: number of days since present lactation
lactation: number of lactations
i79: indicator variable (0 if days $\leq$ 79 and 1 if days > 79)

For the 'current' and 'previous' variables in the milk production data:

- Fit the simple linear regression model with 'current' as the response and 'previous' as the predictor.
- Carry out appropriate tests of hypothesis on the model parameters. Consider the default tests in R and any appropriate additional tests (that can be constructed from the output).
- How could this model be useful to stakeholders and how would you communicate the results and model uncertainty to them?
- Assess the fit of the model. Does anything in your assessment change any of your previous answers?

(b) Multiple regression

For the milk production data, omitting the variable 'i79':

- Carry out a variable selection process and identify the 'best' model.
- Interpret the selected model to determine how it could be used by stakeholders.
- Assess the fit of the selected model and identify any limitations of the model.

The LegumeLegacy project has received funding from the European Union's Horizon 2021 Doctoral Network programme under the Marie Skłodowska-Curie grant agreement No. 101072579

(c) ANOVA modelling

The variables in the survival times data are:

group: variable indicating categories 1 to 12
poison: variable indicating categories 1, 2, 3
treatment: factor indicating levels A, B, C, D
time: survival time in units of 10 hours.

For the survival times data:

- Fit the one-way ANOVA model.
- Fit the two-factor factorial model with an interaction.
- How do these models compare?
- Interpret the two-factor model (with the interaction if needed, without the interaction if not needed) to determine how the results could be used by stakeholders.
- Assess the fit of the model. Based on the assessment, do any of your previous answers need to change?

(d) Jointly modelling continuous and categorical predictors

- Revisit the model selection process from (b) and this time include the categorial variable 'i79'. Assess the fit of the final model and interpret it to determine how it could be used by stakeholders.
- What is the design matrix for the final model?

The LegumeLegacy project has received funding from the European Union's Horizon 2021 Doctoral Network programme under the Marie Skłodowska-Curie grant agreement No. 101072579

SS2.2 – Challenges with Data and Regression Analysis
Lab Session

(a) Revisit the analysis in part (c) of the SS2.1 lab session. Test if transformations of the response in the survival times data would improve the model fit and, if so, re-evaluate the final model.

(b) In the context of (a) the LegumeLegacy multi-site experiment and (b) any individual experiments, what data complexities will need to be taken into consideration for any regression based analyses?
   a. Start by stating the usual assumptions for a regression model, go through each part of the assumptions and come up with examples of where they might not be satisfied.
   b. Then consider and discuss how you might adjust models to address some of these complexities.

The LegumeLegacy project has received funding from the European Union's Horizon 2021 Doctoral Network programme under the Marie Skłodowska-Curie grant agreement No. 101072579

SS2.3 – The design and implementation of experiments and the management of data
Lab Session

(a) Those doing individual experiments:
- Discuss the data you will be collecting and challenges you may face recording data and ensuring data quality.
- Design a data recording template for your individual experiment.

(b) Those managing a site as part of the LegumeLegacy multi-site experiment:
- Discuss the data you will be collecting and challenges you may face recording data and ensuring data quality.
- Examine the LegacyNet data recording template as a starting point for a recording LegumeLegacy data. Discuss any extensions that may be needed.

(c) Those doing multi-site analyses:
- What do you need to consider when collating data from multiple sites?
- To what extent will the database collation system in place for LegacyNet work for LegumeLegacy?
- Examine the LegacyNet data recording template as a starting point for a recording LegumeLegacy data. Discuss any extensions that may be needed.

The LegumeLegacy project has received funding from the European Union's Horizon 2021 Doctoral Network programme under the Marie Skłodowska-Curie grant agreement No. 101072579