10.022 Modelling Uncertainty, 2022

DDW 2D Project Information

Multiple linear regression

- Your goal here is to use multiple linear regression to model a problem that interests you; this is very open-ended; the problem should be related to food security *or* food safety (it does not need to address both).

- To get started, you may consider a simple model with just one or two predictor variables (x variables). During this exploratory stage, you may use r^2 to judge the quality of your model. If the r^2 is low (for instance, lower than 0.25), then you may need to brainstorm or research into other predictor variables. If the r^2 is high, then you are on your way to building a successful model, and you can start fine-tuning your model (such as adding or removing some variables).

  Beware: if the r^2 is extremely close to 1, then chances are something went wrong, e. g. perhaps the Y variable is computed from some x variables in the first place, or perhaps the Y variable is nearly constant for your data.

- To get meaningful analysis and to avoid overfitting, you should use at least 30 rows of data (more would be better). The data could be for 1 single country (but spanning, say, 30 years), or for 30 countries, or a mixture (e. g. 3 countries, 10 years for each country).

  If your data contains missing entries, then you may delete the rows with the missing entries, or try to extrapolate some values for the missing entries. Do not replace the missing entries by an arbitrary value such as 0.

- Your predictor variables (x variables) should be linearly independent from each other, so for instance, avoid situations where $x_2 = 10\,x_1$.

  You are free to choose the x variables based on common sense, literature search, trial and error, data visualization, etc.

- If one of your x variables is a categorical variable with 2 categories (say, 'Asian country' vs 'African country'), then you may represent the categories using 0 and 1 respectively.

  If needed, you may construct separate models for different groups of countries (instead of trying to find 1 model that fits all countries).

- Since *linear* regression is being used here, it would make sense to plot each x variables against Y to check whether the scatter plot *looks* linear. If it is not linear, then some transformation (see Week 2 Class 1) may be required. Depending on the situation, it could even make sense to introduce polynomial terms (e. g. x_1, $(x_1)^2$, $(x_1)^3$), or interaction terms (e. g. x_1, x_2, $x_1 \cdot x_2$), into your model.

Success of your model

- r^2 is not a good measure of how successful a regression is, when comparing models with different numbers of predictor variables. Therefore, after the exploratory stage (see above section), you would need to find a different metric to replace r^2.

  You do not need to invent a new metric, so you may consult Wikipedia, the textbook (Chapter 12), etc. Another hint is to look at the output in *Excel* after you run regression: some entries there you should already understand, such as 'R square', 'Intercept', and 'Lower 95%', but there may be entries that you do not understand. Google those entries; they may help you with finding a new metric, and with determining which x-variables to keep or remove in your model.

- After the exploratory stage, you may end up with multiple 'competing' models, but once you have a suitable metric, then you can use it to determine the 'best' model.
- In your report and *Excel*, you could document and present the following: problem statement, exploratory analysis, data cleaning and transformation (see above section), thought process in finding predictor variables, your chosen metric, steps taken to select a good model, and your final model.