STAT210/410

Assessment 3: Model building, Residual analysis and regression pitfalls

DUE: 13th April 2025, 11:59pm AEDT

Part I

Cadmium exposure in pregnant women can lead to increased risk of birth defects in foetuses (Geng & Wang 2019). Diet is one of the main sources of human exposure to Cadmium with rice identified as one of the main contributors. A study[1] has been undertaken to explore the components in soil which can be used to predict Cadmium accumulation in Rice. 18 soil samples were collected from different rice fields and a number of factors were recorded from each soil sample. The data set, `soil.csv`, contains 12 variables:

- **Soil:** Unique identifier from where the soil was collected. This factor should not be used in the modelling.
- **pH:** the pH of the soil.
- **SOM:** the organic matter in the soil (g/kg)
- **EC:** the electrical conductivity of the soil (ms/cm)
- **Clay:** the amount of clay in the soil (g/kg)
- **Fe:** iron in the soil (g/kg)
- **TN:** total nitrogen in the soil (g/kg)
- **Mn:** MnO content (g/kg)
- **TP:** Total phosphorus (g/kg)
- **CEC:** Cation exchange (cmol/kg)
- **AL:** aluminium in the soil (g/kg)
- **Cd:** the amount of cadmium extracted from the rice plants (mg/kg) which is the response variable.

Question 1 **10 marks**

Use the ggpairs() function to plot the data. In a few sentences, summarise the key correlations between the predictors and the response variable and any correlations between the predictors.

Geng, H. X., & Wang, L. (2019). Cadmium: Toxic effects on placental and embryonic development. *Environmental toxicology and pharmacology*, *67*, 102-107.

[1]Wang, Y., Su, Y., & Lu, S. (2020). Predicting accumulation of Cd in rice (*Oryza sativa L.*) and soil threshold concentration of Cd for rice safe production. Science of the Total Environment, 738, 139805.

NOTE: ggpairs() *is a function within the* GGally *library, so don't forget to load this library before you attempt to produce the plot. To make the correlation text in the plot smaller you can use:*

```
ggpairs(df, columns = ...,
     upper = list(continuous = wrap('cor', size=2)))
```

where df is the name of your data frame, the … specifies the columns you want to choose and you can change 2 to any size that works with your plot panel size.

Question 2 **15 marks**

Fit a main effects (first order) model using all the soil factors (excluding the sample identifier, **Soil**). Use this model to check the four indicators of multicollinearity between your predictors. State which predictors show an indication of multicollinearity. Choose 4 predictors that you think will be useful terms to predict cadmium concentration.

Hint: Look at the correlation values.

Using your chosen predictors, fit a second main effects model and check for multicollinearity.

Question 3 **10 marks**

Fit a complete second order model using the predictors you identified in Question 2. Using the summary output from this model, propose and fit a simpler model. Write the equation for your simplified model.

Question 4 **15 marks**

Run backward stepwise model selection with the complete second order model as the "upper" model. Include all relevant outputs from the stepwise analysis, the summary table for the final model selected through stepwise and the regression equation.

Question 5 **5 marks**

Test the global utility of the simplified model from question 3 and the stepwise model from question 4. Use AIC and the adjusted R^2 to choose which of these two models are the better model to predict cadmium concentration in rice. This will be your final model.

Question 6 **15 marks**

Produce residual plots and carry out residual analysis on the final model to comment on whether the model assumptions have been met. Refer to all five of the residual plots we have talked about in the lectures.

Question 7 **10 marks**

Write a concise, informative conclusion based on your final model from Question 5.

Part II

Rhizospheric soil microbes can have profound effects on plants in Cadmium contaminated soils. Abundance of saprotrophic soil fungi have been found to reduce cadmium accumulation in plant tissues (Cakmak et al., 2023). Wang et al. (2024) conducted an experiment to determine if the soil fungi, Basidiomycota, affected cadmium accumulation in a cadmium hyperaccumulator plant, Black-jack (*Bidens Pilosa*).

The dataset `Cadmium.csv`, contains 3 Variables:

- **Shoot_Cd:** Cadmium concentration in the stems and leaves of *Bidens Pilosa* (mg/kg)
- **Soil_Cd:** Cadmium concentration in the soil (mg/kg)
- **Basid:** Relative abundance of the soil fungus, Basidiomycota (%)

Question 1 **5 marks**

Fit a second order interaction model to the data and check the conditions of the residuals. Which conditions appear to be violated?

Question 2 **10 marks**

Use the `boxcox()` function in R to find suitable transformation/s for your fitted model. Explain why the transformation/s you have chosen is/are deemed appropriate but the other standard power transformations that were discussed in lectures are not.

NOTE: `boxcox()` is a function within the `MASS` library, so don't forget to load this library before you attempt to produce the plot.

Cakmak, I., Lambers, H., Grant, C. A., & Zhao, F. J. (2023). Arbuscular mycorrhizal fungi: key players in avoiding cadmium accumulation in food crops. *Plant and Soil*, *484*(1), 13-32.

Wang, X., Zheng, W. L., Wu, C. L., Han, J. J., Xiang, Y. P., Yang, M. L., He, P., Yu, F. H., & Li, M. H. (2024). Interactive effects of rhizospheric soil microbes and litter on the growth of the invasive hyperaccumulator *Bidens pilosa* in cadmium-contaminated soil. *Frontiers in plant science*, *15*, 1507089. https://doi.org/10.3389/fpls.2024.1507089

Presentation **5 marks**

Marks are allocated for notation and presentation:

- Clear expression, correct use of terminology and mathematical notation
- Presentation of figures and tables; ensure that you include all relevant R output
- Clear and concise annotation of R code: include all R script as an appendix.