Background

A MSc student, Lena, has conducted an experiment measuring corn yields across 10 plots at a site in Bayern, Germany. In addition to yield (kg/ha/yr) she also collected a combination of agronomic, management, and climate data (see ‘ols.reg.data.xlsx’).

She would like to understand the determinants of corn yield in her plots. A key hypothesis of hers is that fertiliser and variety are jointly the main determinant of yields independent of all other agronomic, management, and climate factors.

Problems

To be completed using R file ‘OLS.regression.R’ and data file ‘ols.reg.data.xlsx’

1. Basic regression

Let’s sequentially evaluate the impact of the variables on yields.

First, conduct a basic regression of yield and variety on fertiliser. Call this ‘ols.1’

Second, add in the climate variables, rainfall and temperature. Call this ‘ols.2’

Third, add in the agronomic variables, pH and bulk density. Call this ‘ols.3’

1. Plots

Using ggplot, evaluate the regression coefficients of fertiliser by plotting it against yield (so fertiliser is on x-axis, yield on y-axis). Do this using the two following methods:

1. The ‘geom\_smooth’ method
2. The ‘ab\_line’ method
3. Variable manipulation

Lena suspects that fertiliser and yield may interact non-linearly. She also thinks temperature could be non-linearly related to yield. For temperature she wants to evaluate the impact of temperature changes in relation to the sample mean through ‘mean-standardisation’.

Comparing to the basic regression model (‘ols.1’), run 3 new regressions (‘ols.4’ - ‘ols.6’) as follows:

1. **Polynomial manipulation** (‘ols.4’): Fertiliser enters as the square of fertiliser: fertiliser^2 (polynomial)
2. **Root manipulation** (‘ols.5’): Fertiliser enters as the square root of fertiliser: fertiliser^0.5
3. **Standardisation** (‘ols.6’):Temperature as mean-standardised.
4. Variable interactions

Lena suspects also that the impact of fertiliser might be different depending on how wide the spaces are between rows of corn.

Comparing the basic (‘ols.1’) model with alternatives as follows:

1. **Numeric interaction** (‘ols.7’):Fertiliser and row spacing independently and as a cross-product: *fertiliser* x *row*  spacing (hint: use ‘\*’)
2. **Categoric interaction** (‘ols.8’):Fertiliser and row spacing , distinguishing the effect of fertiliser between ‘small’ from ‘large’ (hint: use ‘:’)
3. Model selection

Based on her analysis, Lena determines the following variables are best for modelling yields in her data. She remains unsure however whether all variables should be included.

Variety

The numeric interactions of row space in cm and fertiliser^0.5

Temperature standardised by ist mean

Rainfall

Bulk density

pH

Using one of two stepwise regression methods, **Forward selection** & **Backwards elimination**, evaluate which model is best. For this you can use one of either R^2, AIC, BIC, F statistic, or p-value, or a combination thereof as evaluation criteria.

What is the best model?