Model Building

Multiple linear regression – Stat 230

In this class, you will explore a modeling-building for multiple linear regression. You will also implement the two new graphical tools introduced in the reading: the component + residual plot and the effect plot.

Discussion of daily prep

Task 1. Take 8 minutes and discuss the answers to the daily prep questions 1-4 with your group.

Modeling the number of species[1]

Johnson and Raven (1973) provide data on the number of species in the Galápagos Islands. There are 30 observations on the following 8 variables:

Variable	Description
Island	Name of the island
Total	Total number of observed species
Native	Number of native species
Area	Area of the island (km^2)
Elev	Elevation of the island (m)
DistNear	Distance to the nearest island (km)
DistSc	Distance to Santa Cruz (km)
AreaNear	Area of the nearest island (km^2)

You can load the data using the command:

```
species <- read.csv("https://aloy.rbind.io/data/galapogos.csv")
```

Ecologists know that number of species on an island is known to be related to the island's area, but are interested in seeing what other variables are also related to the number of species (after

[1]This example is adapted from Exercise 12.20 in *The Statistical Sleuth*. The data come from Johnson, M. P., & Raven, P. H. (1973). Species number and endemism: The galápagos archipelago revisited. Science, 179(4076), 893-895.

accounting for land area). They are also interested in learning whether this differs for native and non-native species.

Task 2. What is the goal of this analysis? How does this inform the model-building process? As you answer these questions, outline the steps you will take during model building and how you will know when you're done.

Task 3. To begin, choose an initial pool of predictor variables. Remember that you will want to look at both univariate and bivariate exploratory plots and summary statistics. Do you need to transform any potential predictors before using them in the model? Are any predictors redundant because they are highly correlated with other predictors?

Task 4. Fit a full regression model using the predictors you identified in the previous task. Does the model reasonable satisfy the assumptions needed for inference? Use component + residual plots to help you check linearity. If not, take remedial action. Are you worried about any influential points?

Task 5. Once you have found a satisfactory full model, conduct inference to answer the researchers questions. Construct effect plots for any variable of primary interest. Write a brief summary of your findings.

Modeling the price of wine[2]

Suppose that you are a wine enthusiast and wish to understand what factors impact the price of fine wine, which would allow you to find good deals. You collect data on 72 wines, recording the following variables:

Variable	Description
wine	Name of the winery (character vector)
price	price (in pounds sterling) of 12 bottles of wine
parker	Robert Parker's rating (out of 100)
coates	Clive Coate's rating (out of 20)
p95	Is Parker's score above 95? 1 = yes, 0 = no
first.growth	Is the wine a first growth? 1 = yes, 0 = no
cult	Is the wine considered to be a cult favorite? 1 = yes, 0 = no
pomerol	Is the wine from Pomerol? 1 = yes, 0 = no
superstar	Is the wine a vintage superstar as awarded by Parker? 1 = yes, 0 = no

To load the data set, run

```
wine <- read.csv("https://aloy.rbind.io/data/bordeaux.csv")
```

Your task is to develop a regression model that enables you to do the following:

- estimate the percentage effect on price of a 1% increase in Parker Points and a 1% increase in Coates Points

[2]This problem was adapted from Sheather (2009),* A Modern Approach to Regression with R*, Springer.

- control for the impacts of first growth, cult favorite wines, and wines from the Pomerol region
- comment on the following claim from Eric Samazeuilh (a courtier):

  > "Parker is the wine writer who matters. Clive Coates is very serious and well respected, but in terms of commercial impact his influence is zero."

- Identify the wines in the data set which, given the values of the predictor variables, are (i) unusually high prices, and (ii) unusually low priced.