STAT202 Assignment 4: Variable selection

David Ewing

Due on 1 pm 21st August

#Introduction

This document contains the analysis for Assignment 3. It explores the aquatic\_toxicity dataset using multiple linear regression to understand the relationships between various predictors and the response variable LC50.

# Step 0: setup

loading libraries:

options(repos = c(CRAN = "https://cloud.r-project.org"))  
  
set.seed(82171165) #set seed   
  
knitr::opts\_chunk$set(  
 echo = TRUE, # Show all code by default  
 message = TRUE, # Include package messages  
 warning = TRUE # Include warnings if they occur  
)  
update.packages(ask = FALSE)

## Warning: package 'foreign' in library 'C:/Program Files/R/R-4.4.1/library' will  
## not be updated

## Warning: package 'MASS' in library 'C:/Program Files/R/R-4.4.1/library' will  
## not be updated

## Warning: package 'nlme' in library 'C:/Program Files/R/R-4.4.1/library' will  
## not be updated

## Warning: package 'survival' in library 'C:/Program Files/R/R-4.4.1/library'  
## will not be updated

library(conflicted)  
library(tidyverse)

## Warning: package 'lubridate' was built under R version 4.4.2

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.4 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2

library(readr)   
library(performance)

## Warning: package 'performance' was built under R version 4.4.2

library(GGally)

## Warning: package 'GGally' was built under R version 4.4.2

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(flextable)

## Warning: package 'flextable' was built under R version 4.4.2

library(broom)

## Warning: package 'broom' was built under R version 4.4.2

library(skimr)

## Warning: package 'skimr' was built under R version 4.4.2

library(data.table)

## Warning: package 'data.table' was built under R version 4.4.2

library(lmtest)

## Warning: package 'lmtest' was built under R version 4.4.2

## Loading required package: zoo  
##   
## Attaching package: 'zoo'  
##   
## The following objects are masked from 'package:data.table':  
##   
## yearmon, yearqtr  
##   
## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

library(leaps)

## Warning: package 'leaps' was built under R version 4.4.2

library(caret)

## Warning: package 'caret' was built under R version 4.4.2

## Loading required package: lattice

library(caTools) #data splitting and reproducibility

## Warning: package 'caTools' was built under R version 4.4.2

library(Metrics)

## Warning: package 'Metrics' was built under R version 4.4.2

conflict\_prefer("filter", "dplyr"); conflict\_prefer("select", "dplyr")

## [conflicted] Will prefer dplyr::filter over any other package.  
## [conflicted] Will prefer dplyr::select over any other package.

# Step 1: Load the dataset

# Load the data  
kungsan <- read.csv("http://stats.apiolaza.net/data/kungsan\_full.csv")  
  
skim\_kungsan <- skim(kungsan) |>  
 select(skim\_variable, n\_missing)  
  
skim\_kungsan

## # A tibble: 4 × 2  
## skim\_variable n\_missing  
## <chr> <int>  
## 1 sex 0  
## 2 height 0  
## 3 weight 0  
## 4 age 0

head(kungsan)

## height weight age sex  
## 1 151.765 47.82561 63 male  
## 2 139.700 36.48581 63 female  
## 3 136.525 31.86484 65 female  
## 4 156.845 53.04191 41 male  
## 5 145.415 41.27687 51 female  
## 6 163.830 62.99259 35 male

# Step 2: Examine the Dataset

set.seed(82171165) # reproducibility  
  
# Create a  
kungsan <- kungsan |>  
 mutate(weight2 = weight^2) |> # new variable `weight2` (weight squared)  
 mutate(sex = factor(sex)) |> # sex factor  
 filter(age >= 12) # Filter 12 years or older  
 my\_kungsan <- kungsan |> sample\_n(100)

summary(kungsan)

## height weight age sex weight2   
## Min. :118.0 Min. :18.26 Min. :12.00 female:221 Min. : 333.3   
## 1st Qu.:146.1 1st Qu.:37.93 1st Qu.:23.00 male :192 1st Qu.:1438.8   
## Median :152.4 Median :43.40 Median :35.00 Median :1883.8   
## Mean :152.0 Mean :42.59 Mean :37.18 Mean :1889.1   
## 3rd Qu.:159.4 3rd Qu.:48.48 3rd Qu.:49.00 3rd Qu.:2350.1   
## Max. :179.1 Max. :62.99 Max. :88.00 Max. :3968.1

summary(my\_kungsan)

## height weight age sex weight2   
## Min. :118.0 Min. :19.62 Min. :12.00 female:58 Min. : 384.9   
## 1st Qu.:147.3 1st Qu.:37.92 1st Qu.:23.75 male :42 1st Qu.:1438.3   
## Median :153.0 Median :43.98 Median :33.50 Median :1934.6   
## Mean :152.7 Mean :42.81 Mean :37.25 Mean :1894.1   
## 3rd Qu.:160.0 3rd Qu.:48.12 3rd Qu.:49.12 3rd Qu.:2315.2   
## Max. :171.1 Max. :58.46 Max. :75.90 Max. :3417.2

#Step 3: Scatterplot Matrix

kung\_splot = ggpairs(my\_kungsan,   
 columns = c("height", "weight", "weight2", "sex"),  
 aes(colour = sex, alpha = 0.7)) +  
 labs(title = "Scatterplot Matrix: Height, Weight, Weight2, and Sex")

kung\_splot

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

