Ecotox Database Handling and Analysis

library(beemixtox1)  
library(tidyverse)  
#> ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──  
#> ✓ ggplot2 3.3.5 ✓ purrr 0.3.4  
#> ✓ tibble 3.1.5 ✓ dplyr 1.0.7  
#> ✓ tidyr 1.1.4 ✓ stringr 1.4.0  
#> ✓ readr 2.0.2 ✓ forcats 0.5.1  
#> ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
#> x dplyr::filter() masks stats::filter()  
#> x dplyr::lag() masks stats::lag()

# Data Pre-processing

## Query from EPA-Ecotox

The following filtering criteria was used to extract the relevant data form the EPA-Ecotox database. In this extraction, the dataset contains 576 CAS numbers.

* **Chemicals**
* **Effect Measurements** 
  + Mortality Group - Mortality
* **Endpoints** 
  + Concentration Based Endpoints
    - LD50
* **Species** 
  + Name(s) / Number(s)
    - honey bee
  + Kingdom: Animals
  + Name Search Type: Common Name
* **Test Conditions** 
  + Test Locations
    - Lab
* **Publication Options**

#Define and set paths  
Dir\_Data = "~/Projects/beemixtox\_public/data-raw/"  
Data.SSD = read.csv(paste0(Dir\_Data,"data/Copy of etc4373-sup-0002-supmat.csv"))  
## usethis::use\_data(Data.SSD) ## The common-name of the EPA Ecotox DB.  
#Load data (USEPA internal not for publication)  
  
#Load data  
  
Data = read.csv(paste0(Dir\_Data,"data/TerrestrialReport.csv"))  
## usethis::use\_data("Data")  
length(unique(Data$CAS.Number))  
#578 CAS

* Selection
  + Life stage: Adult
  + Duration days <=2, restrict to 24 h and 48 h studies.
  + Remove censored LDx.
  + Use CASes with no less than 3 observations. 29 Chemicals
  + Retain only data with units in “AI mg/org”,“AI ng/org”,“AI ug/org”,“mg/bee”,“mg/org”,“ng/org”,“ug/bee”,“ug/org”
  + Exposure type: “Oral via capsule”,“Spray”
* Curation
  + Units transformed to microgram based. (mg, ng ==> ug)

#Subset dataset for relevant parameters  
Data.1 = Data  
Data.1 = Data.1[Data.1$Organism.Lifestage%in%c("Adult"),]  
  
#Subset for relevant exposure times  
Data.1$Observed.Duration..Days.[Data.1$Observed.Duration..Days.=="NR"] = NA  
Data.1$Observed.Duration..Days. = as.numeric(as.character(Data.1$Observed.Duration..Days.))  
  
# Restrict to 24 h & 48 h end-points  
Data.1 = Data.1[Data.1$Observed.Duration..Days.>=1 & Data.1$Observed.Duration..Days.<=2,]  
  
#Restrict to definitive LDx  
Data.1 = Data.1[Data.1$Observed.Response.Mean.Op%in%c(""),]  
  
Data.1 = droplevels(Data.1)  
  
length(unique(Data.1$CAS.Number))  
  
# How many end-points per CAS number?  
NPerCAS = aggregate(Observed.Response.Mean ~ CAS.Number, data =Data.1, length)  
NPerCAS = NPerCAS[order(NPerCAS$Observed.Response.Mean, decreasing = T),]  
  
# Restrict to more than 5 observations  
NPerCAS = NPerCAS[NPerCAS$Observed.Response.Mean>=3,]  
#29 compunts  
  
# Restrict Data.1 for these 29 compounds  
Data.1 = Data.1[Data.1$CAS.Number%in%c(NPerCAS$CAS.Number),]; Data.1 = droplevels(Data.1)  
  
# Homogenize units  
table(Data.1$Observed.Response.Units)  
  
## UNITS  
Data.1$Observed.Response.Mean = as.numeric(as.character(Data.1$Observed.Response.Mean))  
# Retain only useful units  
Data.1 = Data.1[Data.1$Observed.Response.Units%in%c("AI mg/org","AI ng/org","AI ug/org","mg/bee","mg/org","ng/org","ug/bee","ug/org"),]; Data.1 = droplevels(Data.1)  
  
#Transform mg to microgram  
Data.1$Observed.Response.Mean[Data.1$Observed.Response.Units%in%c("AI mg/org","mg/bee","mg/org")] = Data.1$Observed.Response.Mean[Data.1$Observed.Response.Units%in%c("AI mg/org","mg/bee","mg/org")]\*1000 #to ug/bee  
Data.1$Observed.Response.Units[Data.1$Observed.Response.Units%in%c("AI mg/org","mg/bee","mg/org")] = "ug/bee"  
  
#ng to microgram  
Data.1$Observed.Response.Mean[Data.1$Observed.Response.Units%in%c("AI ng/org","ng/org")] = Data.1$Observed.Response.Mean[Data.1$Observed.Response.Units%in%c("AI ng/org","ng/org")]/1000 #to ug/bee  
Data.1$Observed.Response.Units[Data.1$Observed.Response.Units%in%c("AI ng/org","ng/org")] = "ug/bee"  
  
# Homogenize ng/bee  
Data.1$Observed.Response.Units[Data.1$Observed.Response.Units%in%c("AI ug/org","ug/bee","ug/org")] = "ug/bee"  
  
table(Data.1$Observed.Response.Units)  
Data.1 = droplevels(Data.1)  
  
# Exposure media  
table(Data.1$Exposure.Type)  
Data.1 = Data.1[!Data.1$Exposure.Type%in%c("Oral via capsule","Spray"),]; Data.1 = droplevels(Data.1)

## CV calculation

* Restrict the database for chemicals with repeated entries (i.e., more than 3 or 5 LD50 for the same chemical) ==> 16 chemicals

# Approximate variability  
Bee.var.N = aggregate(Observed.Response.Mean ~ CAS.Number, data =Data.1, length)  
Bee.var.Ave = aggregate(Observed.Response.Mean ~ CAS.Number, data =Data.1, mean)  
Bee.var.sd = aggregate(Observed.Response.Mean ~ CAS.Number, data =Data.1, sd)  
  
# Merge  
Bee.var = merge(Bee.var.N,Bee.var.Ave, by="CAS.Number")  
Bee.var = merge(Bee.var,Bee.var.sd, by="CAS.Number")  
names(Bee.var) [c(2:4)] = c("N", "Mean", "sd")  
  
# Restrict to at least 5 cases per CAS  
Bee.var = Bee.var[Bee.var$N>3,] # 16 chemicals  
  
Bee.var$CV.Perc = 100\*(Bee.var$Mean/Bee.var$sd)  
Bee.var = merge(Bee.var, Data.SSD[,c(1,3)], by.x = "CAS.Number", by.y = "CAS.", all.x = T)  
Bee.var = Bee.var[order(Bee.var$N, decreasing = T),]  
  
# Summary of CV%  
summary(Bee.var)  
  
# Check for independency of N and Mean  
plot(Bee.var$N,Bee.var$CV.Perc)  
plot(log10(Bee.var$N),Bee.var$CV.Perc)  
plot(log10(Bee.var$Mean), Bee.var$CV.Perc)  
  
# Which chemicals do we have represented?  
  
#restrict Data.1 to the selected CAS.N  
Data.2 = Data.1  
Data.2 = Data.2[Data.2$CAS.Number%in%Bee.var$CAS.Number,]; Data.2 = droplevels(Data.2)  
Data.2$CAS.Number = as.character(Data.2$CAS.Number)  
  
# Check consistency again  
table(Data.2$Chemical.Grade)  
# Chemical grade is not reported in most cases  
table(Data.2$Organism.Lifestage)  
table(Data.2$Exposure.Type)  
table(Data.2$Number.of.Doses)  
table(Data.2$Observed.Duration..Days.)

aov.1 = aov(Observed.Response.Mean ~ CAS.Number + Conc.1.Type..Author. + Exposure.Type + Observed.Duration..Days., data = Data.2)  
summary(aov.1)  
  
  
# Save reference list only  
Data.2.Ref = unique(Data.2[,c(70,71,72,73,74)])  
##usethis::use\_data(Data.2)  
write.csv(Data.2.Ref, file = "Data.2.Ref\_ECOTOX\_BeeLD50\_Curated.csv", row.names = F)

# Correlations among variables

In this section we

## check the curated dataset

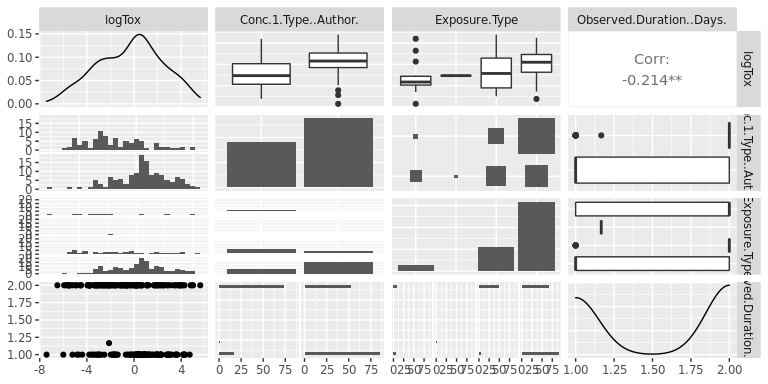
Data.2 <- Data.2 %>% mutate(logTox = log(Observed.Response.Mean))  
mod = lm(log(Observed.Response.Mean) ~ CAS.Number + Conc.1.Type..Author. + Exposure.Type + Observed.Duration..Days., data = Data.2)  
summary(mod)  
#>   
#> Call:  
#> lm(formula = log(Observed.Response.Mean) ~ CAS.Number + Conc.1.Type..Author. +   
#> Exposure.Type + Observed.Duration..Days., data = Data.2)  
#>   
#> Residuals:  
#> Min 1Q Median 3Q Max   
#> -5.6145 -0.7323 -0.0700 0.6656 6.0210   
#>   
#> Coefficients:  
#> Estimate Std. Error t value Pr(>|t|)   
#> (Intercept) 2.1640 0.8232 2.629 0.009238 \*\*   
#> CAS.Number120068373 -7.2625 0.7914 -9.177 < 2e-16 \*\*\*  
#> CAS.Number120928098 -1.9535 0.8967 -2.179 0.030544 \*   
#> CAS.Number121755 -4.6327 0.9124 -5.078 8.76e-07 \*\*\*  
#> CAS.Number1229654663 -4.8739 0.7664 -6.359 1.36e-09 \*\*\*  
#> CAS.Number135410207 -1.1835 0.9040 -1.309 0.191975   
#> CAS.Number138261413 -6.1161 0.6569 -9.311 < 2e-16 \*\*\*  
#> CAS.Number153719234 -6.9559 0.6398 -10.872 < 2e-16 \*\*\*  
#> CAS.Number165252700 -7.6300 0.7372 -10.350 < 2e-16 \*\*\*  
#> CAS.Number183675823 0.8785 0.9099 0.966 0.335454   
#> CAS.Number1910425 0.4239 0.9003 0.471 0.638316   
#> CAS.Number210880925 -7.6220 0.7904 -9.643 < 2e-16 \*\*\*  
#> CAS.Number23135220 -4.0893 0.7982 -5.123 7.08e-07 \*\*\*  
#> CAS.Number50293 -2.1481 0.6586 -3.262 0.001303 \*\*   
#> CAS.Number52315078 -5.6691 0.7336 -7.727 5.29e-13 \*\*\*  
#> CAS.Number52918635 -4.5730 0.8462 -5.404 1.85e-07 \*\*\*  
#> CAS.Number60515 -5.3753 0.8090 -6.644 2.85e-10 \*\*\*  
#> CAS.Number63252 -3.8636 0.7381 -5.234 4.19e-07 \*\*\*  
#> CAS.Number69409945 -4.0208 0.9053 -4.441 1.48e-05 \*\*\*  
#> CAS.Number72208 -3.4981 0.9329 -3.750 0.000232 \*\*\*  
#> CAS.Number736994631 -4.1358 0.9120 -4.535 9.94e-06 \*\*\*  
#> CAS.Number8001352 0.1318 0.8162 0.161 0.871898   
#> CAS.Number86500 -5.1405 0.9329 -5.510 1.10e-07 \*\*\*  
#> CAS.Number91465086 -5.3734 0.8032 -6.690 2.21e-10 \*\*\*  
#> CAS.Number946578003 -5.7006 0.8980 -6.348 1.45e-09 \*\*\*  
#> CAS.Number951659408 -1.7404 0.8059 -2.160 0.031998 \*   
#> Conc.1.Type..Author.Formulation 0.2096 0.2877 0.729 0.467158   
#> Exposure.TypeDiet, unspecified 0.6093 1.5838 0.385 0.700877   
#> Exposure.TypeFood 0.1985 0.4552 0.436 0.663309   
#> Exposure.TypeTopical, general 0.6954 0.4602 1.511 0.132371   
#> Observed.Duration..Days. 0.4129 0.2955 1.397 0.163917   
#> ---  
#> Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
#>   
#> Residual standard error: 1.407 on 199 degrees of freedom  
#> Multiple R-squared: 0.7713, Adjusted R-squared: 0.7369   
#> F-statistic: 22.38 on 30 and 199 DF, p-value: < 2.2e-16  
car::Anova(mod) %>% pander::pander(.)

Anova Table (Type II tests)

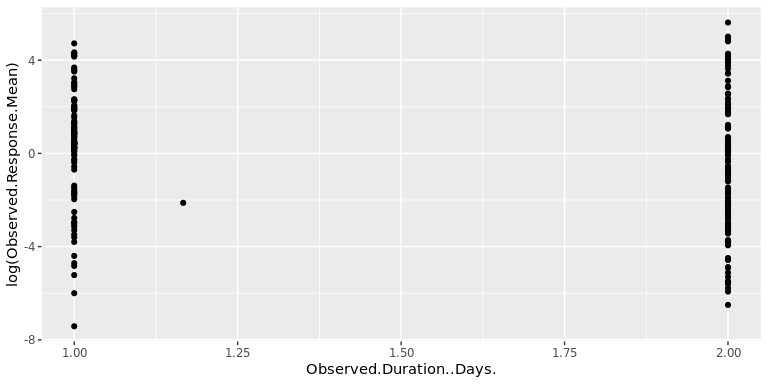
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sum Sq | Df | F value | Pr(>F) |
| **CAS.Number** | 975.7 | 25 | 19.71 | 4.278e-41 |
| **Conc.1.Type..Author.** | 1.051 | 1 | 0.5307 | 0.4672 |
| **Exposure.Type** | 8.464 | 3 | 1.425 | 0.2367 |
| **Observed.Duration..Days.** | 3.866 | 1 | 1.952 | 0.1639 |
| **Residuals** | 394.1 | 199 | NA | NA |

As shown in the Anova Table, none of the covariate variables have a significant effect on the observed response mean.

library(GGally)  
#> Registered S3 method overwritten by 'GGally':  
#> method from   
#> +.gg ggplot2  
ggpairs(Data.2[,c("logTox","Conc.1.Type..Author.","Exposure.Type","Observed.Duration..Days.")])  
#> `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`.  
#> `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



ggplot(data=Data.2,aes(x=Observed.Duration..Days.,y=log(Observed.Response.Mean)))+geom\_point()



ldat <- Data.2 %>% dplyr::select(c(CAS.Number , Conc.1.Type..Author., Exposure.Type,logTox)) %>% pivot\_longer(!logTox,names\_to="variables",values\_to="value")   
ggplot(ldat,aes(x=value,y=logTox))+geom\_boxplot()+geom\_point()+facet\_wrap(~variables,scales = "free")+ theme(axis.text.x = element\_text(angle = 90))

