QSAR model by a GA-MLR regression

Edward Francisco Mendez-Otalvaro, Daniel Alberto Barragan, Isaias Lans-Vargas

2021

Functions to calculate metrics of the model

The metrics that are going to be used are:

- Root mean square error of training set (RMSEC)
- Root mean square error of test set (RMSEP)
- Square of the correlation coefficient of training set (R2tr).
- Square of the correlation coefficient of test set (Also called Q2F2 coefficient).
- Q2F3 coefficient of Todeschini.
- Spearman correlation coefficient of test set (σ).
- Square of the Spearman correlation coefficient of test set ($\sigma 2$).
- Square of the correlation coefficient of cross validation leaving one outside (R2LOO).
- Root mean square error of cross validation leaving one outside (RMSELOO).

All the above metrics are rotinarious in QSAR modelling, in order to get some criteria about a good predictive or ranking model (our main objetive).

```r
Metricas <- function(obs_tr, pred_tr, obs_out, pred_out) {

    ssr_tr <- sum((obs_tr - pred_tr)^2)
    sst_tr <- sum((obs_tr - mean(obs_tr))^2)

    ssr_out <- sum((obs_out - pred_out)^2)
    sst_out <- sum((obs_out - mean(obs_out))^2)

    n_tr <- 20
    n_out <- 7

    ## RMSEC

    RMSEC <- sqrt(ssr_tr/(n_tr))

    ## RMSEP

    RMSEP <- sqrt(ssr_out/(n_out))

    ## R2tr

    R2tr <- (1 - (ssr_tr/sst_tr))

    ## R2out/Q2F2/R2Tropsha

    R2out <- (1 - (ssr_out/sst_out))
    ## Q2F3

    Q2F3 <- (1 - ((ssr_out/n_out)/(sst_tr/n_tr)))

    ## Sigma

    rho <- cor(obs_out, pred_out, method = c("spearman"))

    ## Sigma2

    rho2 <- (rho)^2

    ## Salida

    output <- list(RMSEC = RMSEC, RMSEP = RMSEP, R2tr = R2tr, R2out = R2out, Q2F3 = Q2F3,
        rho = rho, rho2 = rho2)
    return(output)
}

## LOO Metrics

Metrica_LOO <- function(obs, pred) {

    ssr_LOO <- sum((obs - pred)^2)
    sst_LOO <- sum((obs - mean(obs))^2)
    n_LOO <- 20

    ## R2LOO

    R2LOO <- (1 - (ssr_LOO/sst_LOO))

    ## RMSELOO

    RMSELOO <- sqrt(ssr_LOO/n_LOO)

    output <- list(R2LOO = R2LOO, RMSELOO = RMSELOO)
    return(output)
}
```

Importing dataset

Now, the dataset previously prepared is upload (we concatenate the molecular descriptors with the biological activity):

```
Descriptores_I <- read.csv("Descriptores_Alvrunner.csv", stringsAsFactors = FALSE)

Descriptores_II <- read.csv("Descriptores_MOPAC.csv", stringsAsFactors = FALSE)

Descriptores <- cbind(Descriptores_I, Descriptores_II)


## Biological activity

Actividad <- read.csv2("Actividad.csv", stringsAsFactors = FALSE)
```

```r
## Joining molecular descriptors with activity

Dataset <- cbind(Descriptores, Actividad)

## Cleaning labels

Dataset$ID <- NULL
Dataset$IC50..uM. <- NULL

## Renaming biological activity

names(Dataset)[names(Dataset) == "IC50..M."] <- "pIC50"
```

Converting biological activity (EC50) into logarithmic scale, and then, calculating dimension of the dataframe

```r
Dataset[, 2284] <- -log10(Dataset[, 2284])

dim(Dataset)
```

```
## [1]   27 2284
```

So, there are 27 molecules with 2283 molecular descriptors and a biological activity response.

Pretreatment of the dataset (cleaning)

Let's remove NA columns; columns with variance of zero (constant columns) and columns with more than half filled with zeros (Refer to the Manuscript to the criteria selected).

```r
## Removing NA's
Dataset <- Dataset[, !apply(Dataset, 2, function(x) any(is.na(x)))]

## Removing columns with variance equal to zero.
Dataset <- Dataset[, apply(Dataset, 2, var, na.rm = TRUE) != 0]

## Removing columns with constant values
Dataset <- Dataset[, !apply(Dataset, 2, function(x) length(unique(x)) == 1)]

## Removing columns with more than half filled with zeros
Dataset <- Dataset[, colSums(Dataset != 0) > nrow(Dataset)/2]

dim(Dataset)
```

```
## [1]  27 365
```

Since the dataset is very high dimensional `27 X 365`. Let's calculate a correlation matrix for all the descriptors, and then, let's remove the high correlated molecular descriptors (R2 of Pearson >0.99), with this, the multicolinearity between columns could be improved (a problem that could generate bias in our QSAR model)

```r
## Removing high correlated descriptors
library(caret)

## Correlation matrix calculation
Dataset_Cor = cor(Dataset)

## Removing high correlated descriptors with a R2>0.99
```

```r
hc = findCorrelation(Dataset_Cor, cutoff = 0.99)
hc = sort(hc)

## Non correlated descriptor matrix

Dataset_hc = Dataset[, -c(hc)]

## Dimension of the dataset
dim(Dataset_hc)
```

```
## [1]  27 275
```

So now, the dataset has a dimension of `27 X 275`. The improvement of the descriptors is good, but the high dimensionality still appears (n<p, ie; more descriptors than observations)

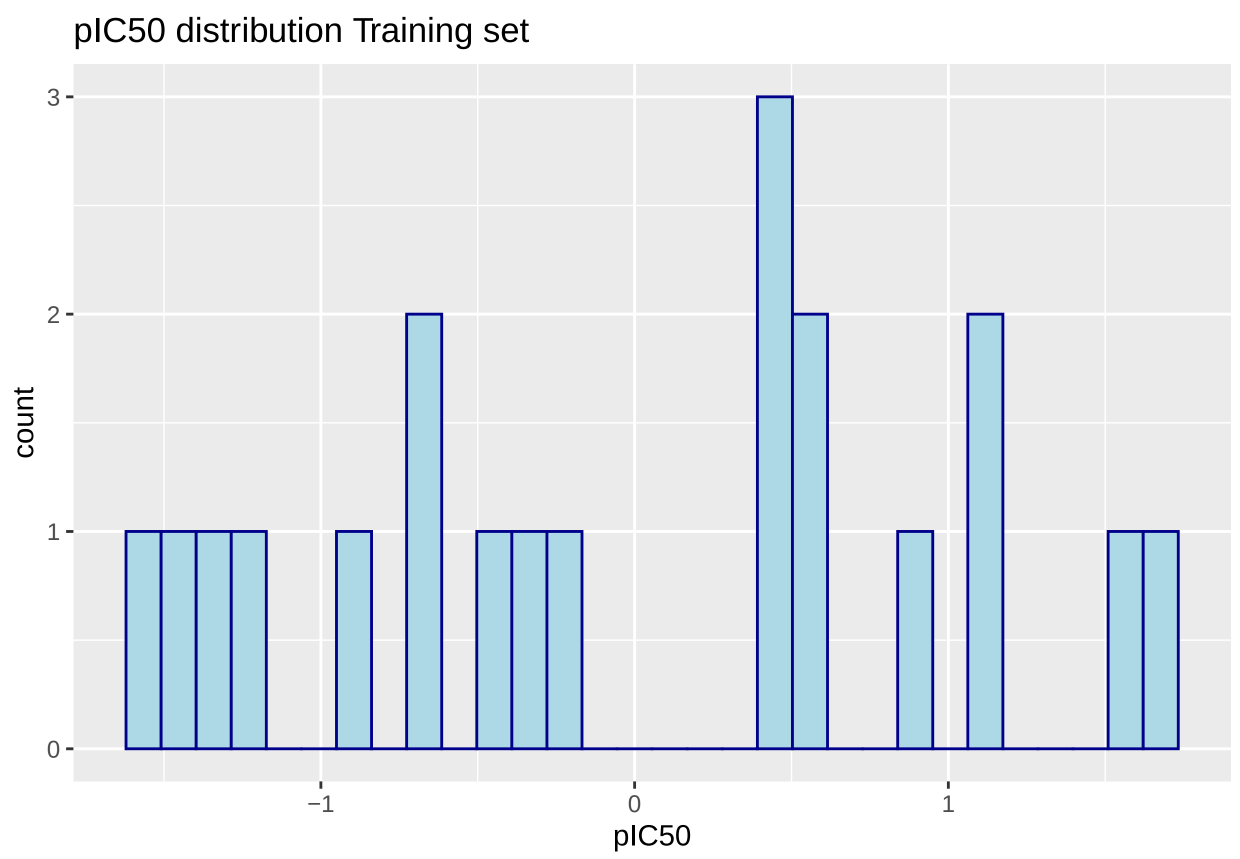
Splitting dataset into training a test set

Since the dataset is very asymmetrical, and there are few observations, let's try to split the dataset in order to get a good balance between both groups (avoiding artifacts by asymetrical splitting). The **caret** function from R allows to carry out this task. Also, the dataset will be scaled subtracting the mean and dividing by standard deviation of data. The ratio will be 70% training and 30% testing.

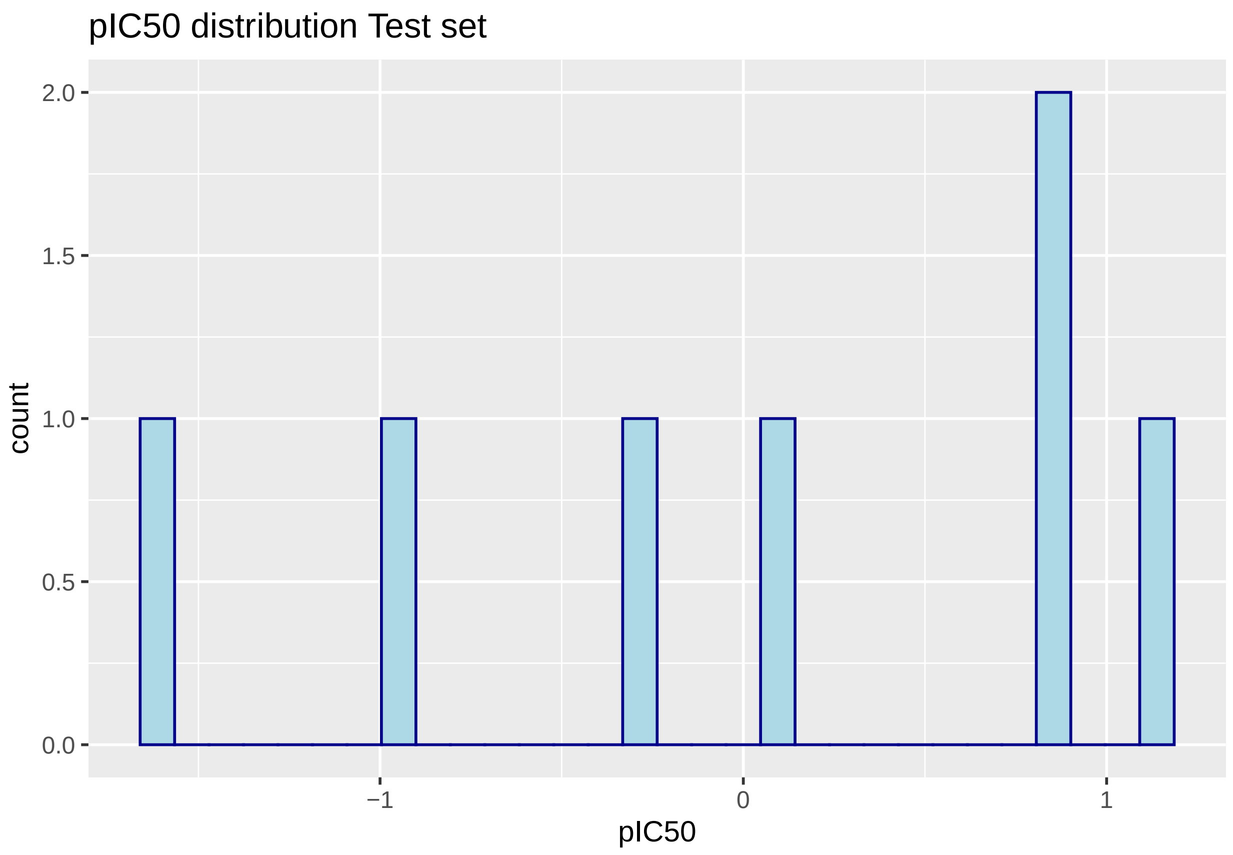
```r
## Scaling
Mean <- apply(Dataset_hc, 2, mean)
SD <- apply(Dataset_hc, 2, sd)

Dataset_S <- as.matrix(scale(Dataset_hc, center = Mean, scale = SD))

centrado <- t(attr(Dataset_S, "scaled:center"))
escalado <- t(attr(Dataset_S, "scaled:scale"))


## Splitting the data in training and test (70% training and 30% test)

set.seed(101)
Muestra <- createDataPartition(Dataset_S[, 275], times = 1, p = 0.7, list = FALSE)

Modelamiento <- as.matrix(Dataset_S[Muestra, ])
Prueba <- as.matrix(Dataset_S[-Muestra, ])
```

Let's apreciate the distribution of the data in the training and test set

```r
library(ggplot2)

## Basic histogram

ggplot(as.data.frame(Modelamiento), aes(x = pIC50)) + geom_histogram(color = "darkblue",
    fill = "lightblue") + ggtitle("pIC50 distribution Training set")
```



```
ggplot(as.data.frame(Prueba), aes(x = pIC50)) + geom_histogram(color = "darkblue",
    fill = "lightblue") + ggtitle("pIC50 distribution Test set")
```



A Student-t statistical test will be carried out to get insight about the differences between groups. The null hypothesis states that the data comes from the same statistical distribution. The alternative hypothesis states that the data does not come from the same statistical distribution. The test will be carried out with a significance level of 95% (p-value = 0.05)

```
t.test(Modelamiento[, 275], Prueba[, 275])
```

```
##
##  Welch Two Sample t-test
##
## data:  Modelamiento[, 275] and Prueba[, 275]
## t = -0.019338, df = 10.475, p-value = 0.9849
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.0032746  0.9859042
## sample estimates:
##    mean of x    mean of y
## -0.002251715  0.006433471
```

Since `p-value>0.05`, then the null hypothesis is accepted, so both data distributions come from the same statistical distribution with a statistical significance level of 95%.

Genetic algorithm for feature selection

Genetic algorithm (GA) is a heuristic technique that mimics natural selection force theory proposed by Darwin. A *population* is defined as a set of candidate solutions and each solution as an *individual*. The value of each solution or *fitness* is calculated and those individuals who have the larger values are selected and combined randomly to produce the next generation (this selection is carrying out like a chromosomal crossing-over and a mutation rate is introduced). The *"reproduction"* process is iterated in an amount of generations, and in each generation is expected that the solution improves.

In feature selection, the individual are a subset of regressors (in this case, the total molecular descriptors) and are encoded as binary (one if the descriptor is included, zero if not); the fitness value is a function that usually corresponds to R2 or RMSE (which is finally the force that exerts selection on the population, i.e., each subset of molecular descriptors adjusted in its own linear regression).

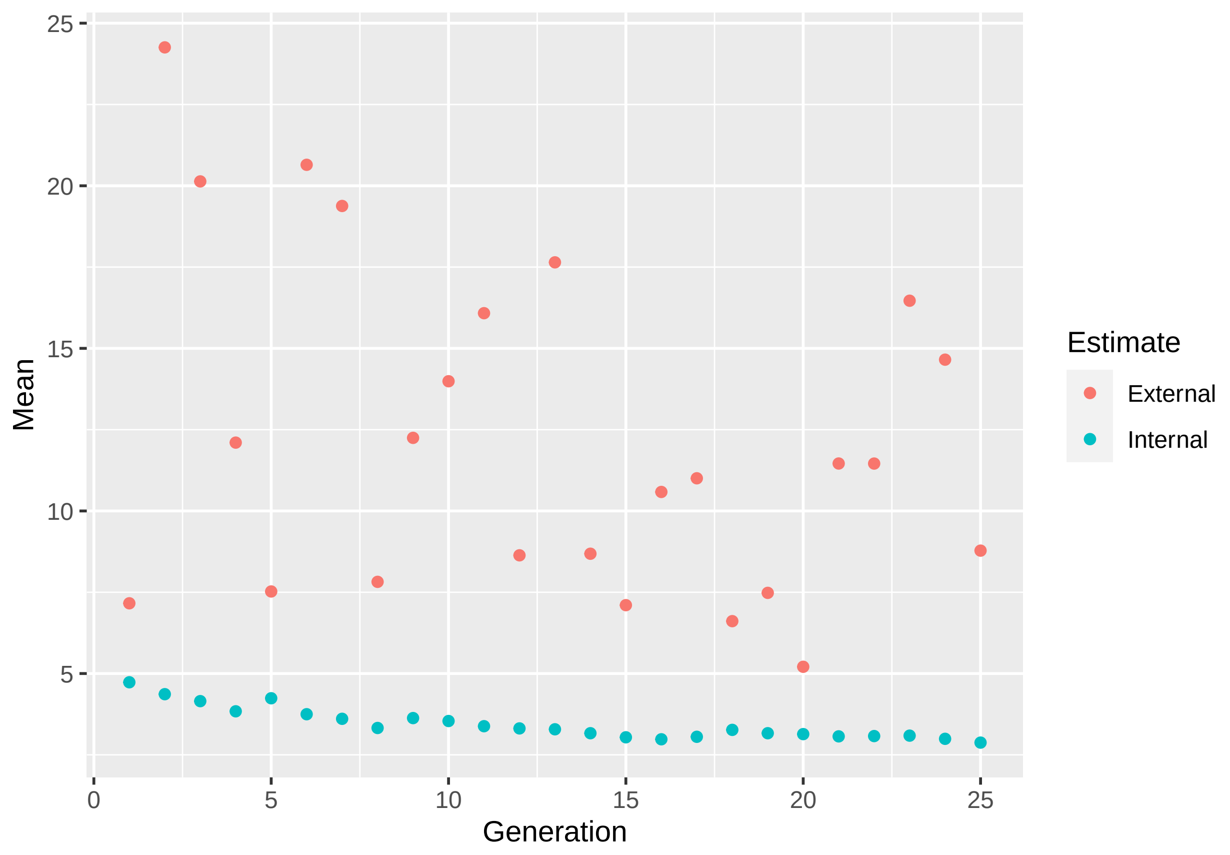
Finally, from all the generations, the best features (regressors or molecular descriptors) are maintained to fit the final linear model with these.

Feature selection by a GA with RMSE as a fitness function and a LOO-CV linear fitting algorithm

```r
## Doing the GA by parallel, since it is very computational resource consuming
library(doParallel)
library(caret)
registerDoParallel(4)  ##Number of cores to parallel
getDoParWorkers()
```

```
## [1] 4
```

```r
## Control parameters for the GA
Control_GA <- gafsControl(functions = caretGA, method = "LOOCV", genParallel = TRUE,
    allowParallel = TRUE, metric = c(internal = "RMSE", external = "RMSE"), maximize = c(internal = FALS
        external = FALSE), verbose = TRUE)

## Adjusting the GA with a CV-LOO linear regression to get best features from the
## dataset

## Since the GA is very time consuming, the code is commented to avoid run it
## again. It takes approximately 6 hours to finish and after that, one can save
## the results (right now, that result is displayed in this document, if the
## reader would like to reproduce that result, please, uncomment this section and
## run it).

# set.seed(100000) GA_lineal <-
# gafs(Modelamiento[,1:274],Modelamiento[,275],iters = 25, gafsControl =
# Control_GA,method = 'lm') GA_lineal View(GA_lineal$optVariables)
# save.image(file='GA_Feature_Selection.RData')
load("GA_Feature_Selection.RData")
plot(GA_lineal)
```



```
GA_lineal
```

```
##
## Genetic Algorithm Feature Selection
##
## 20 samples
## 274 predictors
##
## Maximum generations: 25
## Population per generation: 50
## Crossover probability: 0.8
## Mutation probability: 0.1
## Elitism: 0
##
## Internal performance values: RMSE, Rsquared, MAE
## Subset selection driven to minimize internal RMSE
##
## External performance values: RMSE, Rsquared, MAE
## Best iteration chose by minimizing external RMSE
## External resampling method: Leave-One-Out Cross-Validation
##
## During resampling:
##   * the top 5 selected variables (out of a possible 274):
##     CATS2D_00_LL (75%), CATS2D_08_AL (75%), nR06 (75%), CATS2D_02_LL (70%), CATS2D_04_DA (70%)
##   * on average, 128.5 variables were selected (min = 59, max = 185)
##
```

```
## In the final search using the entire training set:
##    * 174 features selected at iteration 20 including:
##      ALogPS_logP, Sv, Sp, Si, Mv ...
##    * external performance at this iteration is
##
##       RMSE   Rsquared        MAE
##      5.208         NA      5.208
```

Finally, after 25 iterations, the genetic algorithm selected the next five descriptors:

- **CATS2D_00_LL**
- **CATS2D_09_AL**
- **nR06**
- **CATS2D_02_LL**
- **CATS2D_04_DA**

Let's select these molecular descriptors as regressors to fit a classical OLS-MLR model.

Fitting an ordinary least square multiple linear regression model with the selected descriptors by GA

Now, if the five descriptors are selected, it is possible to fit a OLS-MLR model with a system of 20 observations (training set molecules) and 5 regressors (the two molecular descriptors selected by GA). This fit is straightforward and the only assumption is that the square errors from the fitting are normally distributed.

The GA model selected five features from the 274. So the equation looks like (removing CATS word from the features):

$$\mathbf{pIC_{50} = b_0(2D.00.LL) + b_1(2D.09.AL) + b_2(nR06) + b_3(2D.02.LL) + b_4(2D.04.DA) + b_5(Intercept)} \tag{1}$$

(2)

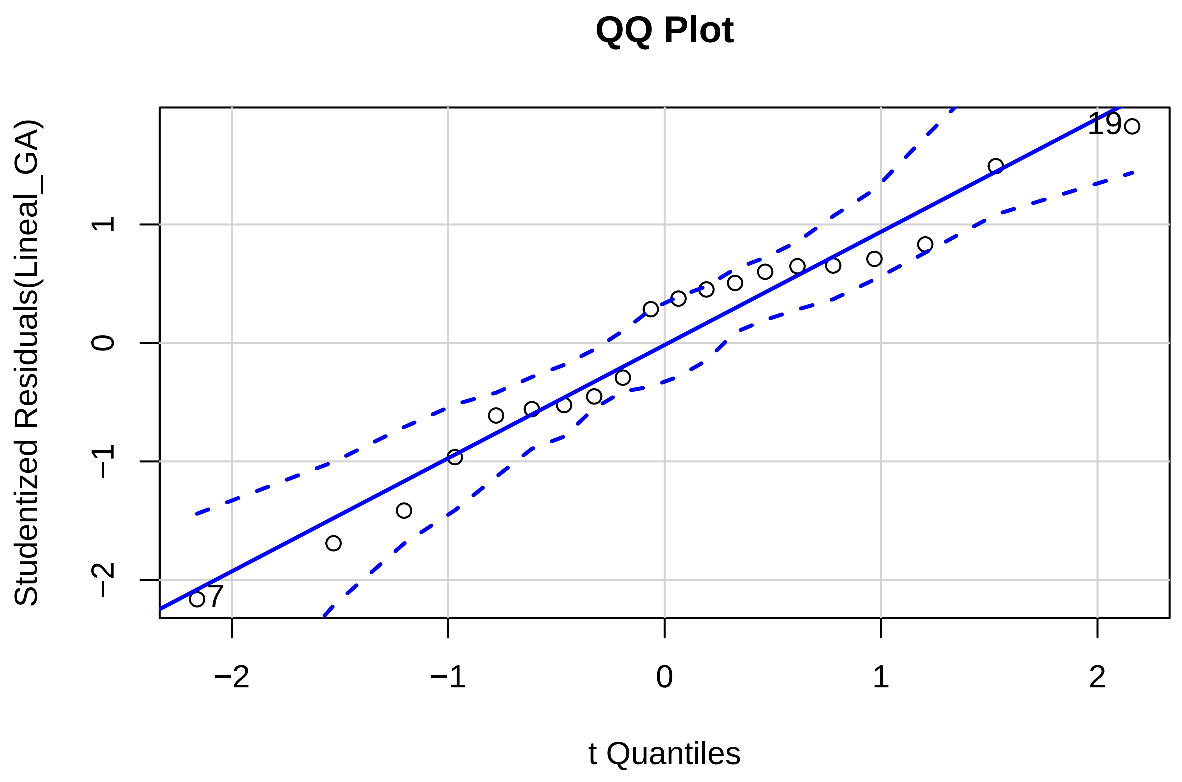
```r
## Selecting only the five descriptors generated by GA
library(dplyr)

Modelamiento_GA <- as.data.frame(Modelamiento)
Modelamiento_GA <- select(Modelamiento_GA, CATS2D_00_LL, CATS2D_08_AL, nR06, CATS2D_02_LL,
    CATS2D_04_DA, pIC50)

Prueba_GA <- as.data.frame(Prueba)
Prueba_GA <- select(Prueba_GA, CATS2D_00_LL, CATS2D_08_AL, nR06, CATS2D_02_LL, CATS2D_04_DA,
    pIC50)

## Fitting the training set into a OLS-MLR:
Lineal_GA <- lm(pIC50 ~ ., data = Modelamiento_GA)
summary(Lineal_GA)
```

```
##
## Call:
## lm(formula = pIC50 ~ ., data = Modelamiento_GA)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3004 -0.3617  0.1872  0.4322  1.1415
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.0374     0.1728  -0.216   0.8317
## CATS2D_00_LL    0.5501     1.3624   0.404   0.6925
## CATS2D_08_AL    0.7096     0.5616   1.264   0.2270
## nR06            0.4647     0.4921   0.944   0.3610
## CATS2D_02_LL   -1.5041     0.6873  -2.188   0.0461 *
## CATS2D_04_DA    0.3626     0.1515   2.394   0.0312 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7434 on 14 degrees of freedom
## Multiple R-squared:  0.6074, Adjusted R-squared:  0.4672
## F-statistic: 4.332 on 5 and 14 DF,  p-value: 0.01364
```

Since this is a typical linear regression fitting, it is possible to get from the output the coefficient values for the regressors, the standard errors, the p-value for statistical significance of each coefficient and the confidence intervals for the coefficients. First, let's plot the square error to check the normality and after that, let's get the confidence intervals for the coefficients.

```
## Getting QQ-Plot (checking squared error normal distribution)
library(car)
qqPlot(Lineal_GA, main = "QQ Plot")
```



```
## [1]  7 19
```

```r
## Confidence intervals for each coefficient

confint(Lineal_GA, level = 0.95)
```

```
##                    2.5 %      97.5 %
## (Intercept)  -0.40793328  0.33313919
## CATS2D_00_LL -2.37193024  3.47211601
## CATS2D_08_AL -0.49487733  1.91412486
## nR06         -0.59076177  1.52021753
## CATS2D_02_LL -2.97808358 -0.03002041
## CATS2D_04_DA  0.03771725  0.68757192
```

Predicting the training dataset with the model (Internal):

```r
Predicho_tr <- predict(Lineal_GA, newdata = Modelamiento_GA)

Observado_tr <- Modelamiento_GA$pIC50
```

Predicting the test dataset with the model (External):

```r
Predicho_out <- predict(Lineal_GA, newdata = Prueba_GA)

Observado_out <- Prueba_GA$pIC50
```

Unscaling data

```r
## Function to unscaling

Desescalar <- function(X) {
    X_desescalada <- (X * escalado[, 275] + centrado[, 275])
}
```

Prediction

```r
## Internal prediction

Predicho_tr <- Desescalar(as.matrix(Predicho_tr))
Observado_tr <- Desescalar(as.matrix(Observado_tr))

Obs_df <- as.data.frame(cbind(Observado_tr, Predicho_tr))
colnames(Obs_df) <- c("Observed", "Predicted")
rownames(Obs_df) <- Muestra[1:20]


## External prediction

Predicho_out <- Desescalar(as.matrix(Predicho_out))
Observado_out <- Desescalar(as.matrix(Observado_out))

Pred_df <- as.data.frame(cbind(Observado_out, Predicho_out))
colnames(Pred_df) <- c("Observed", "Predicted")
rownames(Pred_df) <- c("4", "5", "10", "15", "16", "19", "26")
```

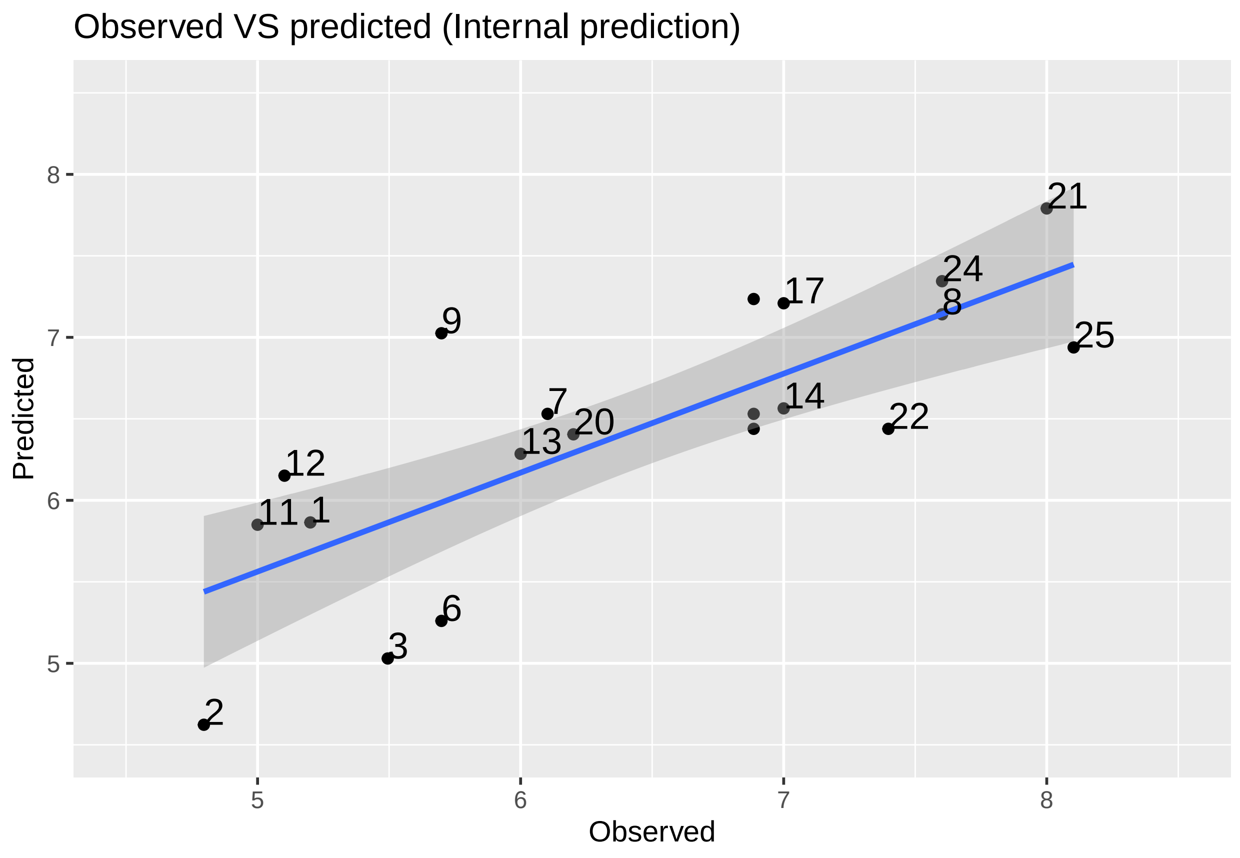
Statistical metrics and plotting of the model

```
Metricas_MLR <- Metricas(Observado_tr, Predicho_tr, Observado_out, Predicho_out)

Metricas_MLR <- (do.call(rbind, Metricas_MLR))
rownames(Metricas_MLR) <- c("RMSEC", "RMSEP", "R2tr", "R2out", "Q2F3", "rho", "rho2")
Metricas_MLR <- round(Metricas_MLR, 1)
print(Metricas_MLR)
```

```
##       [,1]
## RMSEC  0.6
## RMSEP  0.6
## R2tr   0.6
## R2out  0.6
## Q2F3   0.6
## rho    0.8
## rho2   0.7
```

```
library(ggplot2)

# Basic scatter plot Training
ggplot(Obs_df, aes(x = Observed, y = Predicted)) + geom_point() + geom_smooth(method = lm) +
    ggtitle("Observed VS predicted (Internal prediction)") + geom_text(size = 5,
    aes(label = rownames(Obs_df)), hjust = 0, vjust = 0, check_overlap = TRUE) +
    coord_cartesian(xlim = c(4.5, 8.5), ylim = c(4.5, 8.5))
```

```
## `geom_smooth()` using formula 'y ~ x'
```

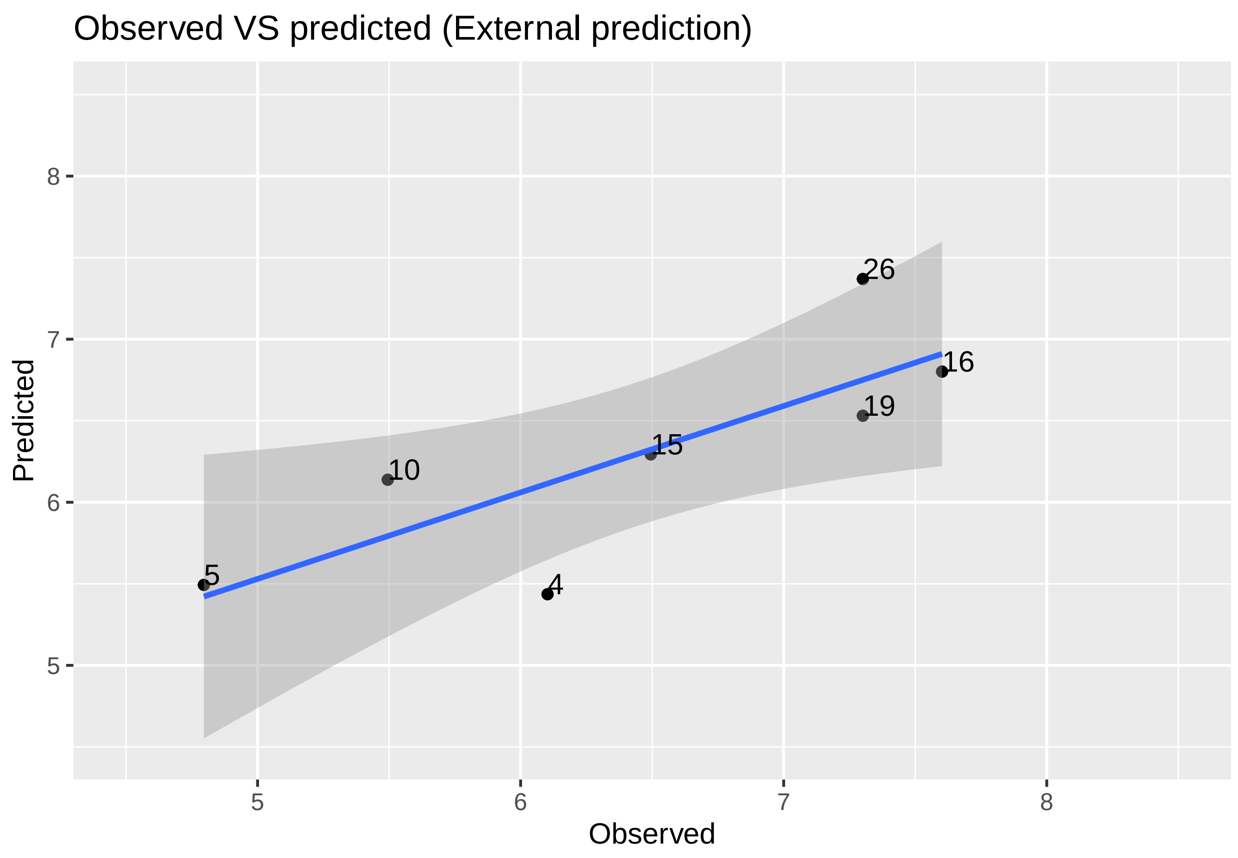


```r
ggsave("gamlrinternal.png")
```

```
## Saving 6.5 x 4.5 in image
## `geom_smooth()` using formula 'y ~ x'
```

```r
# Basic scatter plot Test
ggplot(Pred_df, aes(x = Observed, y = Predicted)) + geom_point() + geom_smooth(method = lm) +
    ggtitle("Observed VS predicted (External prediction)") + geom_text(aes(label = rownames(Pred_df)),
    check_overlap = TRUE, hjust = 0, vjust = 0) + coord_cartesian(xlim = c(4.5, 8.5),
    ylim = c(4.5, 8.5))
```

```
## `geom_smooth()` using formula 'y ~ x'
```



```r
ggsave("gamlrexternal.png")
```

```
## Saving 6.5 x 4.5 in image
## `geom_smooth()` using formula 'y ~ x'
```

Leaving one out cross-validation (LOO-CV) of the data

A statistical metric that is common in QSAR modeling is the RMSE and R2 of the LOO-CV. Let's estimate these parameters

```r
## Storing matrix for the cycle

Pred_CV = matrix()
Modelo_i = list()
```

```r
## Training set

Modelamiento_CV <- as.data.frame(Modelamiento_GA)
set.seed(1e+05)

for (i in 1:nrow(Modelamiento_CV)) {

    ## Removing i row from 1 to the number of rows of the dataset (each cycle left one
    ## out)

    Calibracion_CV <- Modelamiento_CV[-i, ]
    Prueba_CV <- Modelamiento_CV[i, ]

    ## Training the model

    Modelo_GA_CV <- lm(pIC50 ~ ., data = Calibracion_CV)

    ## Predicting the i-row

    Pred_GA_model = predict(Modelo_GA_CV, newdata = Prueba_CV)

    ## Predicted values for each i-row
    Pred_CV[i] <- Pred_GA_model

    ## Model for each i-row
    Modelo_i[[i]] <- Modelo_GA_CV

}
```

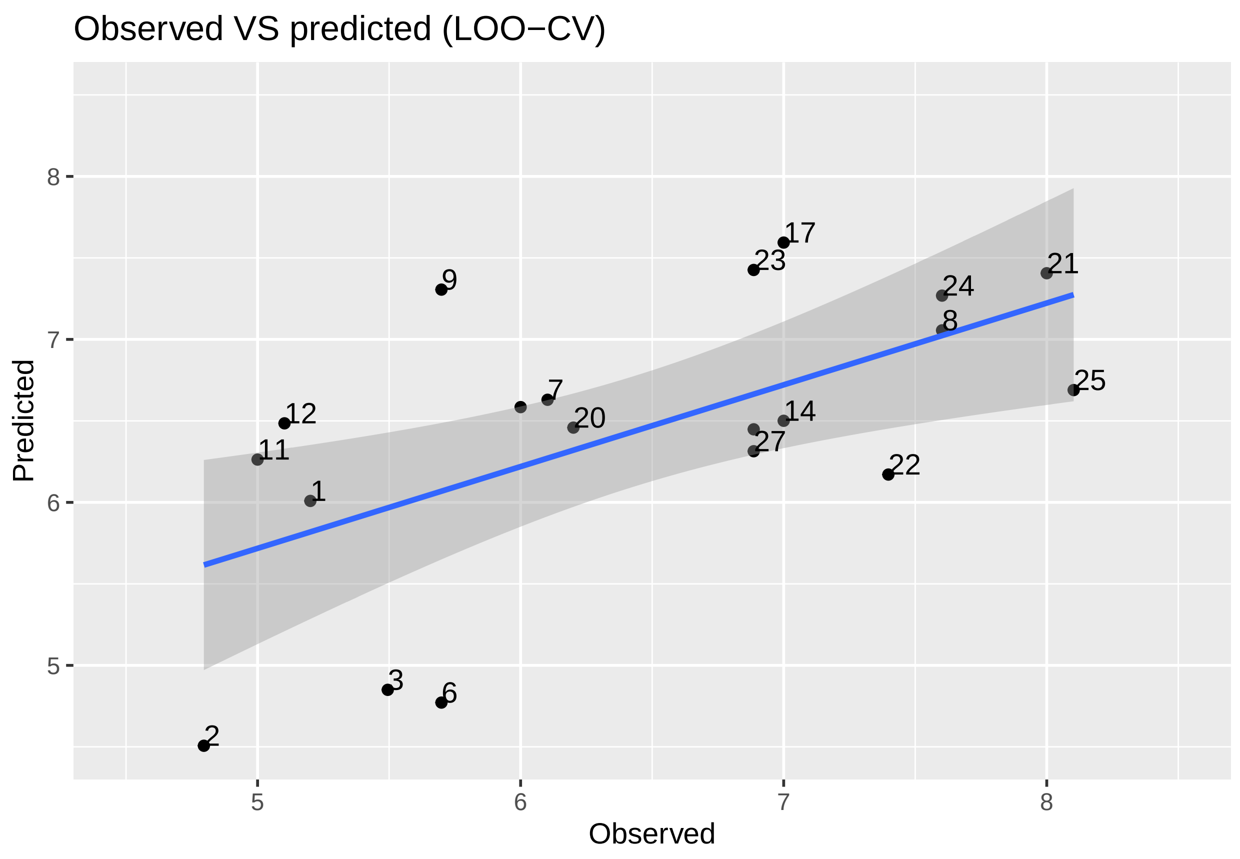
Calculating the metrics and plotting the results

```r
## Unscaling observed and predicted data

Observado_LOO <- Observado_tr
Predicho_LOO <- Desescalar(Pred_CV)

LOO_df <- as.data.frame(cbind(Observado_LOO, Predicho_LOO))
rownames(LOO_df) <- Muestra[1:20]

## Metric calculation
LOO_M <- Metrica_LOO(Observado_LOO, Predicho_LOO)

LOO_M <- (do.call(rbind, LOO_M))
LOO_M <- round(LOO_M, 1)
print(LOO_M)
```

```
##          [,1]
## R2LOO     0.3
## RMSELOO   0.9
```

```r
# Basic scatter plot LOO-CV
ggplot(LOO_df, aes(x = Observado_LOO, y = Predicho_LOO)) + geom_point() + geom_smooth(method = lm) +
    ggtitle("Observed VS predicted (LOO-CV)") + xlab("Observed") + ylab("Predicted") +
    geom_text(aes(label = rownames(LOO_df)), check_overlap = TRUE, hjust = 0, vjust = 0) +
    coord_cartesian(xlim = c(4.5, 8.5), ylim = c(4.5, 8.5))
```

```
## `geom_smooth()` using formula 'y ~ x'
```



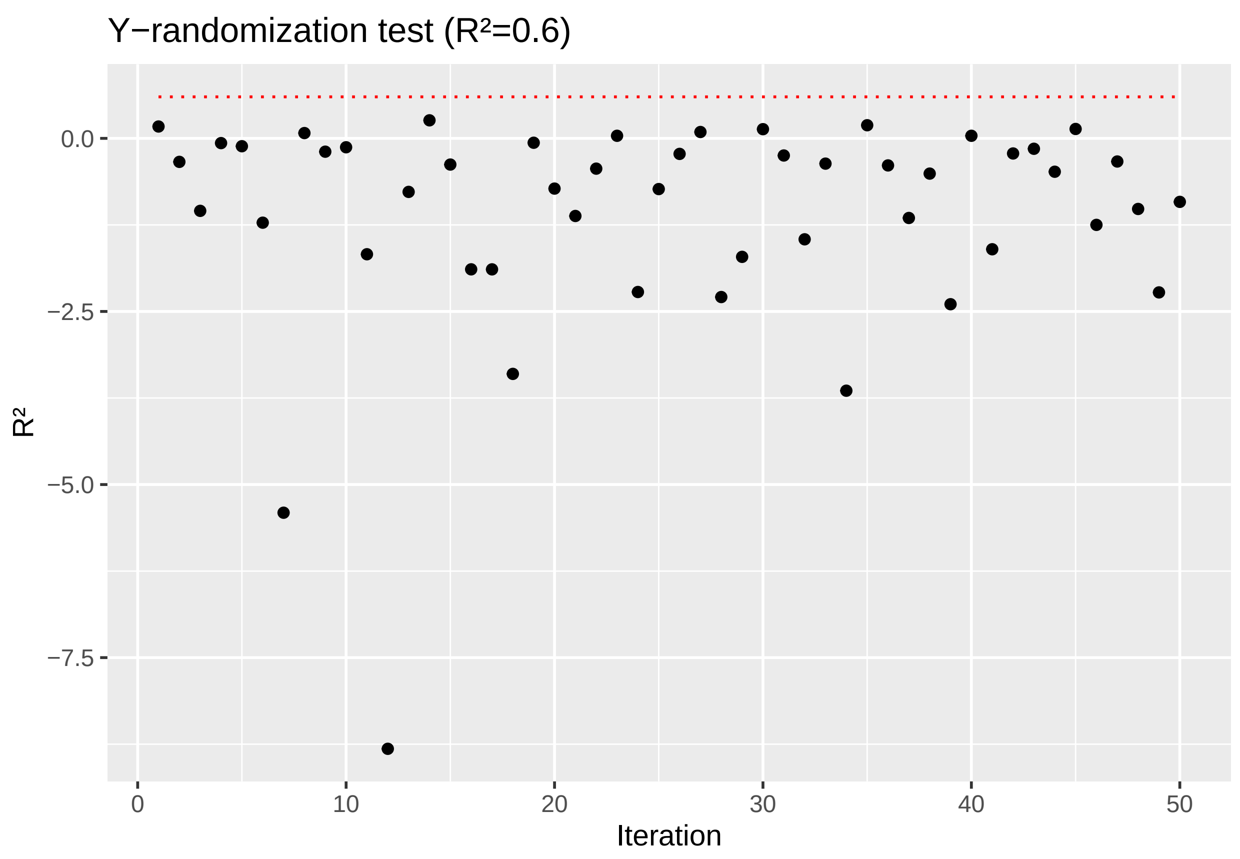
```r
ggsave("gamlrloo.png")
```

```
## Saving 6.5 x 4.5 in image
## `geom_smooth()` using formula 'y ~ x'
```

Y-randomization test for GA-MLR model:

Another statistical test to get insight about the reliability of the QSAR model is the Y-randomization. In this algorithm, the outcome variable (Y=pIC50) is randomized, and in each iteration a new model is build. The R2 for all the random models should be worse than the R2 from the constructed model. This test tries to prove that the model build does not come from random chance.

```r
## 50-fold loop for Y-randomization

## Empty list for store the data that comes from the loop

Score_Y <- list()
set.seed(1e+05)

for (i in 1:50) {

    ## Shuffling the outcome (pIC50), and building training and test sets with this
    ## randomized dataset
    Dataset_Y_shuffle <- as.data.frame(Dataset_S)
    Y_shuffle <- sample(Dataset_S[, 275], replace = FALSE)
    Dataset_Y_shuffle[, 275] <- Y_shuffle
    Modelado_shuffle <- Dataset_Y_shuffle[Muestra, ]
    Prueba_shuffle <- Dataset_Y_shuffle[-Muestra, ]

    Modelado_shuffle <- select(Modelado_shuffle, CATS2D_00_LL, CATS2D_08_AL, nR06,
        CATS2D_02_LL, CATS2D_04_DA, pIC50)
    Prueba_shuffle <- select(Prueba_shuffle, CATS2D_00_LL, CATS2D_08_AL, nR06, CATS2D_02_LL,
        CATS2D_04_DA, pIC50)

    ## Training the model with randomized outcome data

    Modelo_Lineal_Y <- lm(pIC50 ~ ., data = Modelado_shuffle)


    ## Predicting test set with randomized outcome data

    Pred_Y = predict(Modelo_Lineal_Y, newdata = Prueba_shuffle)

    ## Unscaling data

    Observado_Y <- Desescalar(Prueba_shuffle$pIC50)
    Predicho_Y <- Desescalar(Pred_Y)

    ## Model metrics
    Score_Y[[i]] <- Metrica_LOO(Observado_Y, Predicho_Y)
}

## Extracting metrics from the loop
Resul_Y_Random <- as.data.frame(t(unlist(Score_Y)))

par_indexes <- seq(2, 100, 2)
impar_indexes <- seq(1, 99, 2)

RMSEY <- Resul_Y_Random[, par_indexes]
Q2Y <- Resul_Y_Random[, impar_indexes]
Q2Y <- as.data.frame(t(Q2Y))
colnames(Q2Y) <- "R2"
rownames(Q2Y) <- c(1:50)
Q2Y$Iteration <- c(1:50)


library(ggplot2)

# Basic scatter plot Y-randomization
ggplot(Q2Y, aes(x = Iteration, y = R2)) + geom_point() + geom_line(aes(y = 0.6),
    linetype = "dotted", color = "red") + ggtitle("Y-randomization test (R²=0.6)") +
    ylab("R²")
```



```r
ggsave("gamlryrandom.png")
```

```
## Saving 6.5 x 4.5 in image
```

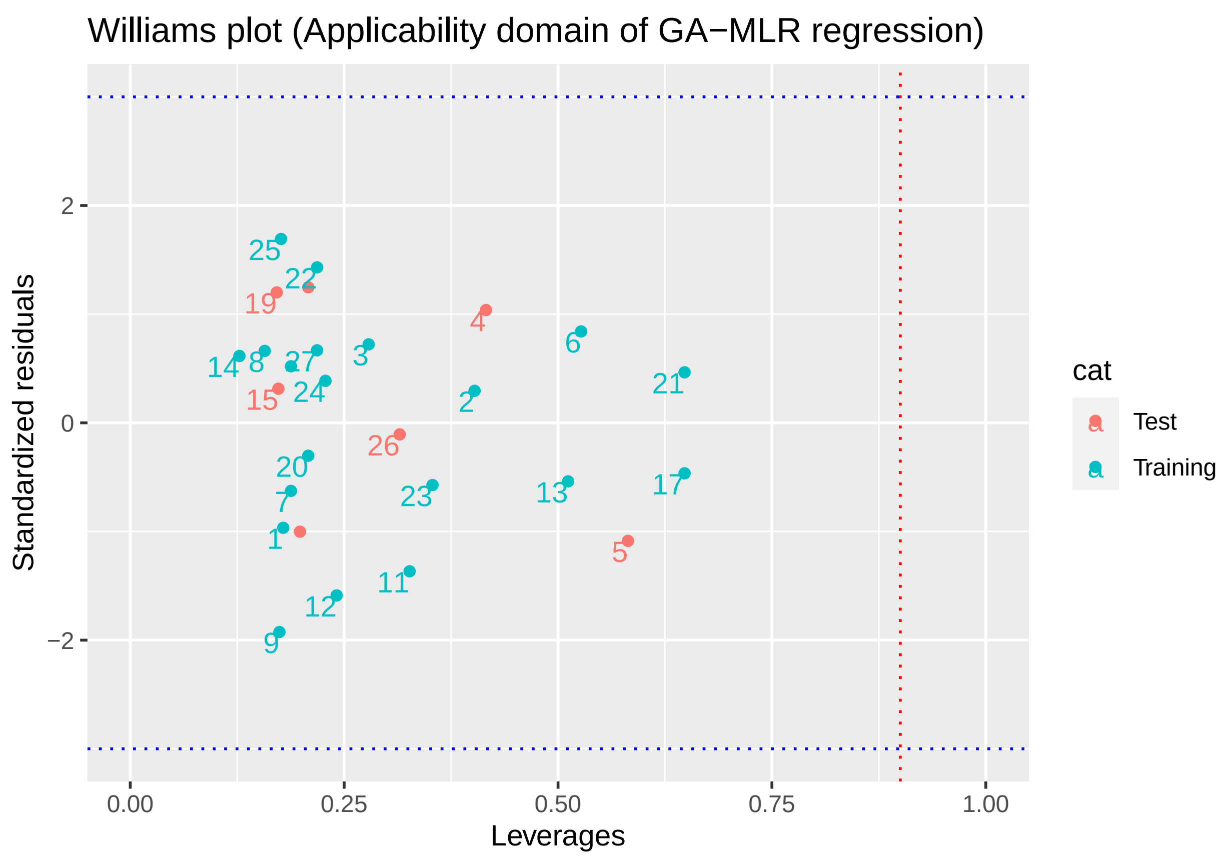
Since none of the randomized models had a better performance than the constructed model (R2=0.6), the QSAR is reliable.

Applicability domain (AD) of the QSAR model

The last step for the development of this QSAR model is the determination of the chemical space gives by the molecular descriptors used to construct the regression. The AD permits to interpolate in a chemical region with reliable results. Finally, the AD also detects possible outliers that are being extrapolated from the model. For more information about AD theory, please refer to the Manuscript.

```r
### For training set

## Getting leverages from the model

hi <- hatvalues(Lineal_GA)  ##Leverages

## Getting standardized residuals from the model

Stand_Residuales <- rstandard(Lineal_GA)

## Estimating warning leverage (every molecule that relies outside this limit,
## can't be interpolated and its predicted value is not trustworthy)
K <- 20
```

```r
h_asterisco <- (3 * (5 + 1))/K  ##Two regressors and 20 observations h*=3(p+1)/n

### For test set

## Getting X matrix from training and testing

X_AD <- as.matrix(Modelamiento_GA)
X_test <- as.matrix(Prueba_GA)

## Mathematical calculation of leverages
H_test <- X_test %*% solve(t(X_AD) %*% X_AD) %*% t(X_test)  ##Hat matrix for test
hi_test <- diag(H_test)  ##Leverages

## Mathematical calculation of standardized residuals from test prediction

obs_test <- Prueba_GA$pIC50
pred_test <- predict(Lineal_GA, newdata = Prueba_GA)

Residuales_test <- (obs_test - pred_test)
Stand_Residuales_test <- Residuales_test/sd(Residuales_test)

## Dataframes for training and test set of leverages and standardized residuals

Will_training <- as.data.frame(cbind(hi, Stand_Residuales))
colnames(Will_training) <- c("Leverages", "Standardized_residuals")
rownames(Will_training) <- Muestra[1:20]

Will_test <- as.data.frame(cbind(hi_test, Stand_Residuales_test))
colnames(Will_test) <- c("Leverages", "Standardized_residuals")
rownames(Will_test) <- c("4", "5", "10", "15", "16", "19", "26")

library(ggplot2)

Will_training$cat <- "Training"
Will_test$cat <- "Test"

ggplot_William <- rbind(Will_training, Will_test)

# Basic scatter plot Williams plot

ggplot(ggplot_William, aes(x = Leverages, y = Standardized_residuals, color = cat)) +
    geom_point() + xlim(0, 1) + geom_text(label = rownames(ggplot_William), check_overlap = TRUE,
    hjust = 1, vjust = 1) + geom_vline(xintercept = 0.9, linetype = "dotted", color = "red") +
    geom_hline(yintercept = c(-3, 3), linetype = "dotted", color = "blue") + ggtitle("Williams plot (Ap
    ylab("Standardized residuals")
```



```r
ggsave("gamlrad.png")
```

With the Williams plot, the AD of the model is determined, and apparently all the training and test molecules are correctly interpolated into the chemical space defined by the molecular descriptors selected.

Final parameters of the linear model:

```r
library(knitr)

Mol_Descriptor <- c("CATS2D_00_LL", "CATS2D_08_AL", "nR06", "CATS2D_02_LL", "CATS2D_04_DA",
    "Intercept")
Coefficient <- as.numeric(c(0.5501, 0.7096, 0.4647, -1.5041, 0.3626, -0.0374))
CImin <- as.numeric(c(-2.37193024, -0.49487733, -0.59076177, -2.97808358, 0.03771725,
    -0.40793328))
CImax <- as.numeric(c(3.47211601, 1.91412486, 1.52021753, -0.03002041, 0.68757192,
    0.33313919))
Error <- as.numeric(c(1.3624, 0.5616, 0.4921, 0.6873, 0.1515, 0.1728))

Final_parameters_Lineal <- as.data.frame(cbind(Error, CImax, CImin, Coefficient))

Final_parameters_Lineal <- cbind(Mol_Descriptor, Final_parameters_Lineal)

round(Final_parameters_Lineal[, 2:5], 3)
```

```
##   Error  CImax  CImin Coefficient
## 1 1.362  3.472 -2.372       0.550
## 2 0.562  1.914 -0.495       0.710
## 3 0.492  1.520 -0.591       0.465
## 4 0.687 -0.030 -2.978      -1.504
## 5 0.152  0.688  0.038       0.363
## 6 0.173  0.333 -0.408      -0.037
```

```
kable(Final_parameters_Lineal)
```

Mol_Descriptor	Error	CImax	CImin	Coefficient
CATS2D_00_LL	1.3624	3.4721160	-2.3719302	0.5501
CATS2D_08_AL	0.5616	1.9141249	-0.4948773	0.7096
nR06	0.4921	1.5202175	-0.5907618	0.4647
CATS2D_02_LL	0.6873	-0.0300204	-2.9780836	-1.5041
CATS2D_04_DA	0.1515	0.6875719	0.0377172	0.3626
Intercept	0.1728	0.3331392	-0.4079333	-0.0374

Now, the QSAR model is finished.

Virtual screening

After the development of the QSAR model; two databases of molecules were taken from **PubChem** and **ZINC**. This molecules were filter by two parameters: **Glucosamine-like molecules** and **Lipinski rule of five**.

The *Lipinski rule of five* is a empirical rule for druglikeness of molecules, the physicochemical parameters for this rule are:

- No more than 5 hydrogen bond donors.
- No more than 10 hydrogen bond acceptors.
- A molecular mass less than 500 g/mol.
- An octanol-water partition coefficient (Log P) that does not exceed 5.

The *glucosamine-like molecules* are mandatory to kept into the databases since the QSAR was development under glucosamine derivative molecules, and the chemical space is defined or interpolated under this scaffold.

After the filtering we got 155 unique molecules in SMILES format. The correct protonation state for each molecule was carried out by **Gypsum-DL software** with a pH between 7.0 and 7.2. The tautomer and isomer distribution for each molecule was carried out by the same software, and for each molecule was generated two (310) and three (465) possible tautomer/conformer and protonated states in order to get a better insight of all the chemical states on the dataset.

Finally, the datasets were uploaded into OCHEM web platform in order to calculate mechanistic interpretable 2D and 3D molecular descriptors (the same molecular descriptors used for the QSAR development).

Two variants

```
## Invoking datasets
library(dplyr)

## Alvadescriptors
alvdesc_2 <- read.csv("alvdesc-2.csv", stringsAsFactors = FALSE)

## MOPAC
mopac_2 <- read.csv("mopac-2.csv", stringsAsFactors = FALSE)
```

```r
## Joining both molecular descriptor datasets

mopac_2$SMILES <- NULL

VS_db <- cbind(alvdesc_2, mopac_2)
SMILES <- VS_db$SMILES
```

Cleaning datasets

```r
## Converting to numeric
VS_db <- apply(VS_db, 2, as.numeric)
VS_db <- as.data.frame(VS_db)
VS_db$SMILES <- NULL

## Scaling data
Mean_VS <- as.data.frame(t(Mean))
SD_VS <- as.data.frame(t(SD))
Mean_VS$pIC50 <- NULL
SD_VS$pIC50 <- NULL

VS_db <- select(VS_db, names(Mean_VS))

VS_db <- as.matrix(VS_db)
VS_db <- scale(VS_db, center = Mean_VS, scale = SD_VS)
```

Predicting activites

```r
Predicho_out_VS <- as.data.frame(predict(Lineal_GA, newdata = as.data.frame(VS_db)))

Predicho_out_VS <- Desescalar(Predicho_out_VS)
ID <- c(1:310)

Predicho_out_VS <- cbind(ID, Predicho_out_VS)

colnames(Predicho_out_VS) <- c("Index", "pIC50")

## Prediction
Predicho_out_VS <- as.data.frame(Predicho_out_VS)
class(Predicho_out_VS)
```

```
## [1] "data.frame"
```

```r
write.csv(Predicho_out_VS[order(Predicho_out_VS$pIC50, decreasing = TRUE), ], file = "VS_GAMLR-2.csv",
    row.names = FALSE)
head(Predicho_out_VS)
```

```
##   Index    pIC50
## 1     1 6.834702
## 2     2 6.834702
## 3     3 4.634541
## 4     4 4.634541
## 5     5 9.059868
## 6     6 9.059868
```

Three variants

```r
## Invoking datasets
library(dplyr)

## Alvadescriptors
alvdesc_3 <- read.csv("alvdesc-3.csv", stringsAsFactors = FALSE)

## MOPAC
mopac_3 <- read.csv("mopac-3.csv", stringsAsFactors = FALSE)

## Joining both molecular descriptor datasets

mopac_3$SMILES <- NULL

VS_db3 <- cbind(alvdesc_3, mopac_3)
SMILES3 <- VS_db3$SMILES
```

Cleaning datasets

```r
## Converting to numeric
VS_db3 <- apply(VS_db3, 2, as.numeric)
VS_db3 <- as.data.frame(VS_db3)
VS_db3$SMILES <- NULL

## Scaling data
Mean_VS <- as.data.frame(t(Mean))
SD_VS <- as.data.frame(t(SD))
Mean_VS$pIC50 <- NULL
SD_VS$pIC50 <- NULL

VS_db3 <- select(VS_db3, names(Mean_VS))

VS_db3 <- as.matrix(VS_db3)
VS_db3 <- scale(VS_db3, center = Mean_VS, scale = SD_VS)
```

Predicting activites

```r
Predicho_out_VS3 <- as.data.frame(predict(Lineal_GA, newdata = as.data.frame(VS_db3)))

Predicho_out_VS3 <- Desescalar(Predicho_out_VS3)
ID <- c(1:465)

Predicho_out_VS3 <- cbind(ID, Predicho_out_VS3)

colnames(Predicho_out_VS3) <- c("Index", "pIC50")

## Prediction
Predicho_out_VS3 <- as.data.frame(Predicho_out_VS3)
class(Predicho_out_VS3)
```

```
## [1] "data.frame"
```

```
write.csv(Predicho_out_VS3[order(Predicho_out_VS3$pIC50, decreasing = TRUE), ], file = "VS_GAMLR-3.csv"
    row.names = FALSE)
head(Predicho_out_VS3)
```

```
##   Index    pIC50
## 1     1 6.834702
## 2     2 5.448862
## 3     3 5.448862
## 4     4 4.634541
## 5     5 4.634541
## 6     6 4.634541
```