assignment_2

2024-10-16

```
library(caTools)
```

```
## Warning: package 'caTools' was built under R version 4.3.3
```

```
library(ggplot2)
```

Problem 1. Regression

```
data <- read.csv("qsar_aquatic_toxicity.csv", sep = ";", header = FALSE)
names(data) <- c(
    "TPSA",
    "SAacc",
    "H050",
    "MLOGP",
    "RDCHI",
    "GATS1p",
    "nN",
    "C040",
    "LC50"
)

head(data)
```

```
##     TPSA   SAacc H050 MLOGP RDCHI GATS1p nN C040  LC50
## 1   0.00   0.000    0 2.419 1.225  0.667  0    0 3.740
## 2   0.00   0.000    0 2.638 1.401  0.632  0    0 4.330
## 3   9.23  11.000    0 5.799 2.930  0.486  0    0 7.019
## 4   9.23  11.000    0 5.453 2.887  0.495  0    0 6.723
## 5   9.23  11.000    0 4.068 2.758  0.695  0    0 5.979
## 6 215.34 327.629    3 0.189 4.677  1.333  0    4 6.064
```

a. Split the data into a training and test set

```
# Use 70% of dataset as training set and remaining 30% as testing set
sample <- sample.split(data$LC50, SplitRatio = 0.7)
train  <- subset(data, sample == TRUE)
test   <- subset(data, sample == FALSE)
```

```r
dim(train)
```

```
## [1] 382   9
```

```r
dim(test)
```

```
## [1] 164   9
```

```r
# Fit linear regression model on training data
model <- lm(LC50 ~ ., data=train)

summary(model)
```

```
##
## Call:
## lm(formula = LC50 ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4792 -0.7715 -0.1126  0.5888  4.9593
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.745166   0.315497   8.701  < 2e-16 ***
## TPSA         0.024997   0.003184   7.851 4.41e-14 ***
## SAacc       -0.014653   0.002543  -5.762 1.74e-08 ***
## H050        -0.003053   0.074602  -0.041 0.967377
## MLOGP        0.394378   0.077961   5.059 6.64e-07 ***
## RDCHI        0.586729   0.168072   3.491 0.000539 ***
## GATS1p      -0.577951   0.192562  -3.001 0.002868 **
## nN          -0.165961   0.063152  -2.628 0.008944 **
## C040         0.020611   0.097915   0.211 0.833391
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.244 on 373 degrees of freedom
## Multiple R-squared:  0.4633, Adjusted R-squared:  0.4518
## F-statistic: 40.26 on 8 and 373 DF,  p-value: < 2.2e-16
```

```r
# Predict on training and test datasets
pred_train <- predict(model, newdata=train)
pred_test <- predict(model, newdata=test)
```

```r
# Adding predictions columns to the datasets
train$predicted_LC50 <- pred_train
test$predicted_LC50 <- pred_test
```

```r
# Evaluate model: calculate MSE, RMSE, and R-squared for training and test sets
mse_train <- mean((train$LC50 - train$predicted_LC50)^2)
rmse_train <- sqrt(mse_train)
r2_train <- 1 - (sum((train$LC50 - train$predicted_LC50)^2) / sum((train$LC50 - mean(train$LC50))^2))
```

```r
mse_test <- mean((test$LC50 - test$predicted_LC50)^2)
rmse_test <- sqrt(mse_test)
r2_test <- 1 - (sum((test$LC50 - test$predicted_LC50)^2) / sum((test$LC50 - mean(test$LC50))^2))
```

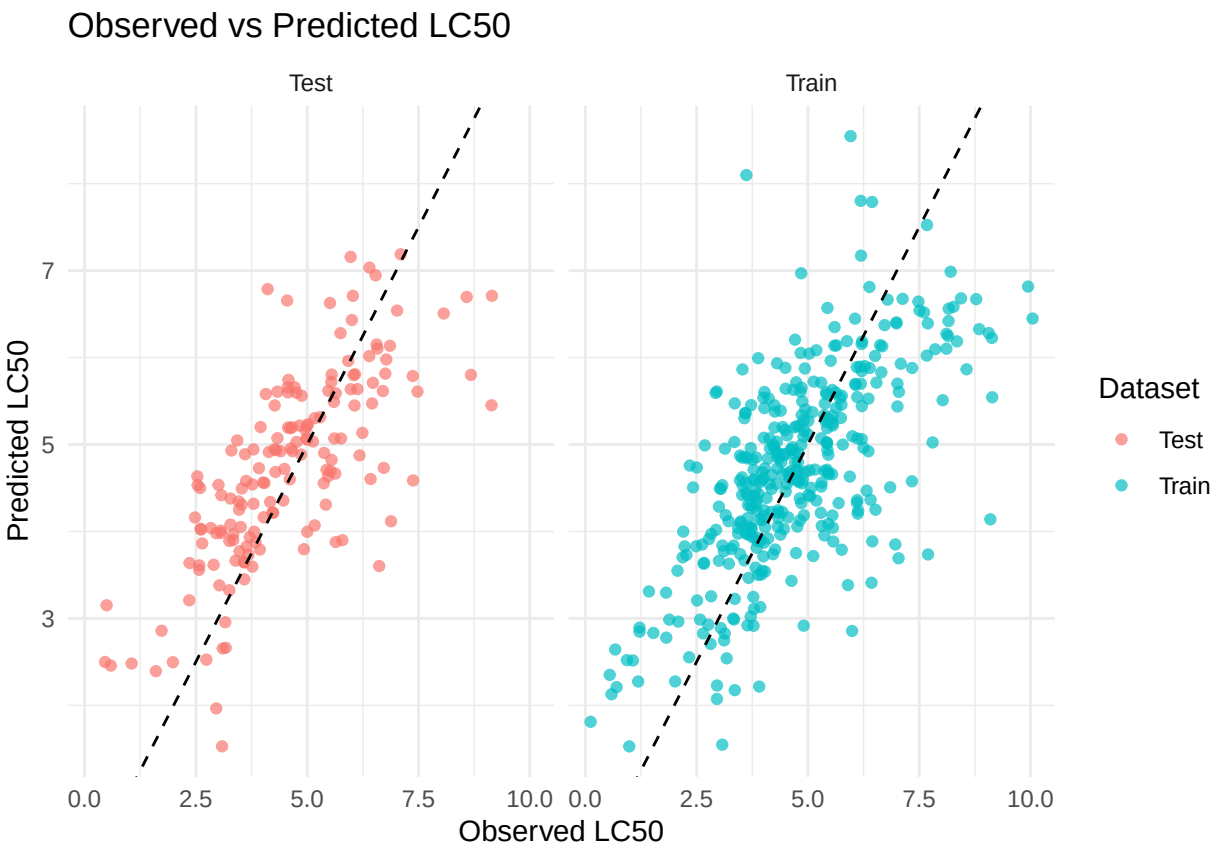
```r
cat(paste0(
  "Training Metrics:\n",
  "MSE (Train): ", mse_train, "\n",
  "RMSE (Train): ", rmse_train, "\n",
  "R-squared (Train): ", r2_train, "\n\n",

  "Test Metrics:\n",
  "MSE (Test): ", mse_test, "\n",
  "RMSE (Test): ", rmse_test, "\n",
  "R-squared (Test): ", r2_test, "\n"
))
```

```
## Training Metrics:
## MSE (Train): 1.5117735325755
## RMSE (Train): 1.22954200114331
## R-squared (Train): 0.46334553654679
##
## Test Metrics:
## MSE (Test): 1.24628014266055
## RMSE (Test): 1.1163691784802
## R-squared (Test): 0.528552778281214
```

```r
# Combine data for plotting
train$Dataset <- 'Train'
test$Dataset <- 'Test'
plot_data <- rbind(train, test)

# Plot observed vs predicted LC50 values
ggplot(plot_data, aes(x = LC50, y = predicted_LC50, color = Dataset)) +
  geom_point(alpha = 0.7) +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed") +
  labs(title = "Observed vs Predicted LC50", x = "Observed LC50", y = "Predicted LC50") +
  theme_minimal() +
  facet_wrap(~Dataset)
```


Observed vs Predicted LC50
Test
Train
Predicted LC50
Observed LC50
7
5
3
0.0
2.5
5.0
7.5
10.0
Dataset
Test
Train