Finding MLR Models for the Octanol/Water Partial Coefficients

Cicily Balachandar

cicily.balachandar@uconn.edu

Abstract

Five different models will be examined, where the predictor variables of interest are HBA1, HBA2, HBD, PSA, and MR. General linear hypothesis testing is used to see if these variables contribute to the model and if they can be useful towards predicting the octanol-water partial coefficient of a molecular compound.

Multicollinearity will be examined and modifications to the models are made if necessary. Then, coefficients of determination and AICs help in selecting the "best" model. The model is validated using the testing portion of data that is separate from the training data used to select the model.

Section 1. Introduction

The goal here is to use MLR methods to predict the partition coefficient between water and logPn of certain compounds, based on the number of structures present within those compounds. This is important because being able to model structures of a compound for the purpose of estimating certain measurable properties can be very useful.

In the original analysis of the data, from the training set of data, multiple regression models were constructed. In particular, three models were examined. One model only used the count of structural features of the molecular compounds. The second model also included additional predictors related to hydrogen bonds. The final model added in polar surface area (PSA) and molar refractivity (MR) as predictors. Then, k-fold cross validation was used on the three models.

To make this analysis unique, I will be testing more than three models. More specifically, I will be testing whether HBA1 and HBA2 both need to be included in model 2, or if one of them can be dropped or combined. Also, instead of having model 3 with both PSA and MR, I will separate them into two models, with one model having PSA and the other having MR. Then another model will include both PSA and MR.

Section 2. Description of Data

There are two sets of data, one being the training data and the other being the test data. The training data includes 87 observations, which are 87 different molecular compounds. The test data has 40 observations, which are 40 molecular compounds. The predictor variables are the counts of each type of structural feature found in the molecular compounds, which comes to 19 predictors. The other predictor variables include HBA1 (sum of lone pairs on atoms), HBA2 (sum of acceptor atoms), HBD (count of hydrogen bond donor atoms), PSA (polar surface area),

and MR (molar refractivity). The Y variable is the $\log P_N$ of each molecular compound, also known as the n-octanal-water partition coefficient.

Section 3. Methods and Models

The assumptions are that the predictor variables are independent from each other, the compounds are independent from each other, and the random error is distributed as normal with a mean zero and a finite variance. Also, the error variance must be constant. MLR regression will be the main method used here.

There will be 5 models in total. Model (1) contains only the 23 structural features of the compounds. Model (2) contains the structural feature predictors plus HBA1, HBA2, and HBD. Model (3) has the structural feature predictors, HBA1, HBA2, HBD, and PSA. Model (4) includes the has the structural feature predictors, HBA1, HBA2, HBD, and MR.

(1) $Y = B0 + B1*X1 + B2*X2 + B3*X3 + B4*X4 + \ldots + B23*X23 + \varepsilon$

(2) $Y = B0 + B1*X1 + B2*X2 + B3*X3 + B4*X4 + \ldots + B23*X23 + B24*X24 + B25*X25 + B26*X26 + \varepsilon$

(3) $Y = B0 + B1*X1 + B2*X2 + B3*X3 + B4*X4 + \ldots + B23*X23 + B24*X24 + B25*X25 + B26*X26 + B27*X27 + \varepsilon$

(4) $Y = B0 + B1*X1 + B2*X2 + B3*X3 + B4*X4 + \ldots + B23*X23 + B24*X24 + B25*X25 + B26*X26 + B28*X28 + \varepsilon$

And the final full model is as follows:

(5) $Y = B0 + B1*X1 + B2*X2 + B3*X3 + B4*X4 + \ldots + B28*X28 + \varepsilon$

In Model (5), all of the predictor variables are included in the model, which include the 23 structural features of the compounds, HBA1, HBA2, HBD, PSA, MR.

Section 4. Analysis of Data

Before fitting the models, we make sure that the assumptions are satisfied.

To check for the normality of the error terms, the Shapiro-Wilk test was used on each of the 5 models.

Table 1: Shapiro-Wilk Tests for Normality of Error Terms

Model	p-value
1	0.03116
2	0.08911
3	0.07264
4	0.1116
5	0.05916

The null hypothesis is that the error terms are normal. Thus, any model with a p-value greater than a 5% significance level satisfies the normality condition. From Table 1, all models except Model 1 have normality of the error term.

Next, we test the constancy of variance assumption for each of the models.

Table 2: Test for Heteroscedasticity

Model	p-value
1	0.9874
2	0.81424
3	0.77019
4	0.78768
5	0.78408

The null hypothesis for this test is that the variance is constant. Thus, the models with p-values greater than a significance level of 5% satisfy this assumption. Based on Table 2, all of the models satisfy the constant variance assumption.

Then, the Durbin-Watson test is used to check for serial correlation.

Table 3: Durbin-Watson Test

Model	D-W Statistic
1	2.099783
2	2.067411
3	2.184309
4	2.109908
5	2.162276

If the D-W statistic is close to 2, then it is assumed that there is no first-order serial correlation present, and the errors are assumed independent. From Table 3, all of the D-W statistics are close to 2 and thus, there is no problem with serial correlation.

After checking the assumptions, the data is transformed to make sure the normality of error terms condition is satisfied. The Yeo-Johnson transformation was used. Furthermore, this transformation will be applied to all of the models, so they can be compared to each other on an equal basis. The Shapiro-Wilk test is conducted again.

Table 4: Shapiro-Wilk Test of the Transformed Data

Model	p-value
1	0.1393
2	0.2739
3	0.2786
4	0.3306
5	0.2442

From Table 4 above, all of the p-values are greater than 0.05, and thus the normality condition for the error terms is satisfied for all models.

Now that the necessary changes to the data are made, the previous assumptions are checked again and the predictor variables and models can be examined.

First, general linear hypothesis testing will be used to see if HBA1, HBA2, and HBD are needed in the model. Below, are the hypotheses that will be tested:

H0: B24 = B25 = B26 = 0

H1: At least one of B24, B25, or B26 is not equal to 0

Using an F test here, the corresponding F statistic is 3.5917 and the p-value is 0.01827. Based on a significance level of 5%, we reject H0 because the p-value is less than 0.05. Thus, HBA1, HBA2, and HBD are needed in the model.

Some other tests are whether PSA is useful to the model, MR is useful to the model, and PSA and MR are useful to the model. A significance level of 0.017 will be used for the 3 tests.

For testing PSA, the following hypotheses will be used:

H0: B27 = 0

H1: B27 ≠ 0

The corresponding F statistic for this test is 8.435, and the p-value is 0.005069. Since the p-value is less than 0.017, the null hypothesis is rejected and PSA is useful to the model.

For testing MR, the following hypotheses are used:

H0: B28 = 0

H1: B28 ≠ 0

The corresponding F statistic is 0.5381 and its p-value is 0.466. Clearly, the p-value is greater than 0.017 and we fail to reject the null hypothesis. Thus, MR is not a very useful predictor for the model.

For testing both PSA and MR, the hypotheses are as follows:

H0: B27 = B28 = 0

H1: Either B27 or B28 ≠ 0

The F statistic is 4.2853 and the p-value is 0.01806. The p-value is greater than 0.017, so we fail to reject H0 and including both PSA and MR is not useful to the model.

Table 5 shows that Models 3 and 5 have the highest R^2's, which means more of the variation in the response is explained by the predictor variables compared to the other models.

Table 5: R^2 for Each Model

Model	R^2
1	0.792
2	0.822
3	0.843
4	0.8235
5	0.8436

Looking at Model 5, multicollinearity seems to be an issue. Table 6 below indicates the predictor variables with VIF's greater than 10.

Table 6: VIF's of Predictor Variables

Predictor Variable	VIF
RSO2NR	20.878116
RSR	10.066463
C	732.170167
RINGS	15.024296
HBA1	60.357484
HBA2	216.404862
HBD	27.758193
PSA	165.922126
MR	1142.010428

Table 6 shows that there is a severe multicollinearity problem. To fix this problem 6 new models will be considered. A new variable X23 = HB was created, which combines HBA1, HBA2, and HBD. In addition, another variable X21 = CARO was created, which combines C and AROMATIC.

Model 7: $Y = B0 + B1*X1 + B2*X2 + B3*X3 + B4*X4 + \ldots + B20*X20 + \varepsilon$

Model 7 has all of the predictor variables except for RINGS, C, AROMATIC, HB, PSA, and MR.

Model 8: $Y = B0 + B1*X1 + B2*X2 + B3*X3 + B4*X4 + \ldots + B20*X20 + B23*X23 + \varepsilon$

Model 8 is the same as Model 7, except HB is included in the model and HB is also standardized.

Model 9: $Y = B0 + B1*X1 + B2*X2 + B3*X3 + B4*X4 + \ldots + B20*X20 + B21*X21 + \varepsilon$

Model 9 is the same as Model 7, but instead this model also includes the variable CARO.

Model 10: $Y = B0 + B1*X1 + B2*X2 + B3*X3 + B4*X4 + \ldots + B20*X20 + B24*X24 + \varepsilon$

Model 10 has all of the predictor variables in Model 7, except for RSO2NR, to avoid collinearity since PSA is now added to the model.

Model 11: $Y = B0 + B1*X1 + B2*X2 + B3*X3 + B4*X4 + \ldots + B20*X20 + B25*X25 + \varepsilon$

Model 11 is the same as Model 7, but it now includes MR.

Model 12: $Y = B0 + B1*X1 + B2*X2 + B3*X3 + B4*X4 + \ldots + B20*X20 + B22*X22 + \varepsilon$

Model 12 is the same as Model 7, but the predictor, RINGS, is added.

These models were derived based on the correlation values found between certain predictor variables that were moderate to high in strength. These values are summarized in Table 7 below.

Table 7: Correlation coefficients between certain predictor variables

C	MR	0.9833149
RINGS	MR	0.9097226
PSA	MR	0.7962643
C	PSA	0.6950086
RINGS	PSA	0.7135984
AROMATIC	MR	0.7937571
AROMATIC	PSA	0.6325589
C	AROMATIC	0.796384
PSA	RSO2NR	0.796384
HB	MR	0.8769566
HB	PSA	0.9627081
HB	AROMATIC	0.6540786

Based on the above table, certain predictor variables were able to be combined or dropped to create the new models.

It is then found that Models 7, 9, 10, and 12 all have VIFS less than 5 for each of their predictor variables. Thus, there is no issue of multicollinearity for those models.

Model 8 has a VIF of 8.013446 for the standardized HB, which means that the model with HB may not be as safe to use. And Model 11 has a VIF of 5.120225 for MR, which indicates a possible problem with multicollinearity.

Next, the coefficients of determination and MSE's can be found for the models.

Table 8: R^2 and MSE's of the New Models

Model	R^2	MSE
7	0.5789	1.2465
8	0.5053	1.5630
9	0.7512	0.786
10	0.4659	1.6636

11	0.7328	0.844
12	0.5789	1.3305

All of the assumptions for these new models have been checked. Thus, from Table 8, we can see that Models 9 and 11 have higher coefficients of determination compared to the other models.

Table 9: AIC and SBC for Models 9 and 11

Model	AIC	SBC
9	-5.096952	39.28939
11	1.119413	45.50576

Based on Table 9, the model with smaller values for AIC and for SBC are generally preferred. Thus, Model 9 seems to be the best model out of the other models examined here.

Now, Model 9 will be validated by testing it on a separate set of data.

The coefficient of determination for the testing data set is 0.8928.

The MSPR is 0.2508163. From Table 9, we see that the MSE is 0.786. The MSPR is less than the MSE. It looks like Model 9 has good predictive ability.

Section 5. Concluding Remarks

In the original 5 models, we found that most of the assumptions were satisfied except for the normality assumption in Model 1. Thus, the Yeo-Johnson transformation was applied to the model. Then, the conclusions we reached from the general linear hypothesis testing was that HBA1, HBA2, and HBD were useful in the model. We also found that PSA and MR were useful for predicting the response of $\log P_N$. The coefficients of determination were found for the models and they were high for all of the models. Although, when examining the VIFS, there seemed to be severe multicollinearity issues present. Thus, changes to the models needed to be made.

To fix the multicollinearity issue, 6 new models were devised. In general, it was found that even when combining HBA1, HBA2, and HBD into a new variable and standardizing it, the VIF was greater than 10. Thus, this predictor variables (which are related to characteristics of hydrogen) are too closely related to the other predictors and should be excluded from the model. There seemed to be no other severe multicollinearity issues with the other modified models.

Then, after examining the coefficients of determination and MSE's of the new models, we found that Model 9 and Model 11 seemed to be the best in terms of explaining the variability in $\log P_N$. Based on the AICs, it was concluded that Model 9 seemed to be the best.

In general, overlooking multicollinearity can be easy, but then the models will not be good for inference and hypothesis testing. With many predictor variables it can hard to fix for multicollinearity. The easiest way would be to combine predictor variables that have similar characteristics and to standardize them. If the issue is still there, then it might be helpful to examine different models with different combinations of the predictor variables causing the high VIFs to find models without severe multicollinearity issues.

References

Lopez, K., Pinheiro, S., & Zamora, W. J. (2021, July 12). *Multiple linear regression models for predicting the N-octanol/water partition coefficients in the Sampl7 Blind Challenge - Journal of Computer-aided molecular design*. SpringerLink. Retrieved March 26, 2022, from https://link.springer.com/article/10.1007/s10822-021-00409-2