QSAR Lab Report

Donna Wrublewski

August 24, 2019

# Purpose

# Introduction

One focus of Chemical Informatics is to try to predict the behavior of a compound based on its structure. Especially, behavior in a biological system can be very complex and difficult to predict in a simple way. Some examples are:

* drug activity
* drug side effects
* environmental toxicity

These are complex types of behavior and are difficult to predict. One way to approach this problem is known as a QSAR – Quantitative Structure-Activity Relationship. Peter et al. (2019) The most common approach to QSAR is rather like constructing a Calibration Curve, except that multiple variables are used in building the model. You start with a series of “descriptors” - numbers that describe the structure of the molecule which are easy to compute or measure, for example: MW, solubility, number of hydrogen- bond-donors, dipole moment, and many others. Using these it is possible to build a multiple regression model: an equation to predict the activity of interest:

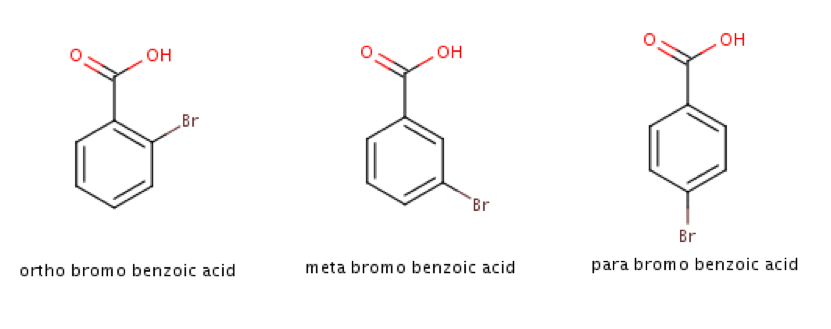
activity = intercept + (const1)\*(descriptor1) + ( const2)\*(descriptor2) + (const3)\*(descriptor3) ...

In this case, the “activity” represents the Dependent variable (also called the Response variable) and the “descriptors” represent multiple Independent variables (also called Explanatory variables). This equation is an expansion of the standard formula for a straight line:

y = mx + b

in this case “b” is the intercept and “m” is the slope.

The sample data file for this exercise (benzoate-pka.txt) gives data on a series of substituted benzoic acids. The Hammett plot has been used in physical organic chemistry to understand the effect of electron-donating and electron-withdrawing groups on a given reaction. The Hammett sigma value could be considered an electronic molecular descriptor, although it is not one typically used in cheminformatics applications. This works well if all of your data points have the same substitution pattern - either all meta or all para. What if you have all three substitution patterns present in your dataset?



Substituted benzoic acids.

We will use this method to model the pKa values for a series of substituted benzoic acids. Substituents on the benzene can affect the pKa of benzoic acid – electron withdrawing substituents lower the pKa (more acidic), and electron donating substituents increase the pKa (less acidic). The **Sigma value** of a substituent is a measure of how strongly electron-donating or electron-withdrawing it is. But the value of Sigma, and the pKa both depend on its location on the benzene ring: *ortho*, *meta* or *para*. As a result, two different versions of sigma are included in the data file: **sigma.m** for when the substituent is meta, and **sigma.p** for when the substituent is para. For *ortho* substituents, the sigma.p value can be used.

As a further refinement, we can add Indicator Variables to the model. In the dataset there are three indicator variables that are either one or zero to indicate the position of the substituent in each molecule in the dataset, whether it is ortho, meta or para.  
There are three other descriptors as well: TPSA describes the total polar surface area of the molecule, logP describes its solubility in water versus in octanol, and MR is the molar refractivity which has to do with the polarizability of the molecule.

# Results

## Exploratory Analysis of *pKa* values

### pKa Values

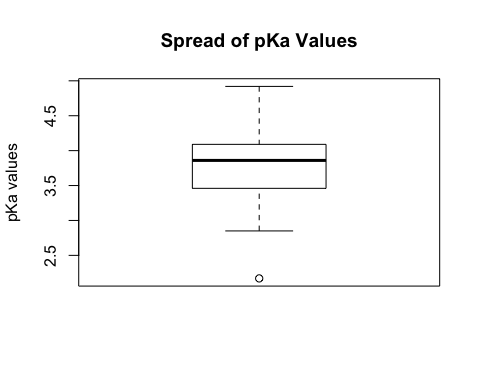
There are 33 values and the median value is 3.86.

Raw pKa Data

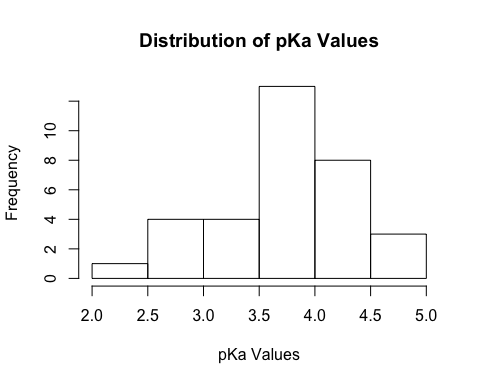
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | pka | sigma.m | sigma.p | meta | ortho | para | tpsa | logp | mr |
| H | 4.20 | 0.00 | 0.00 | 0 | 0 | 0 | 37.30 | 1.38480 | 33.40130 |
| O-CH3 | 3.91 | -0.07 | -0.17 | 0 | 1 | 0 | 37.30 | 1.69320 | 38.36730 |
| O-Ph | 3.46 | 0.06 | -0.01 | 0 | 1 | 0 | 37.30 | 3.05180 | 58.83730 |
| O-OCH3 | 4.09 | 0.12 | -0.27 | 0 | 1 | 0 | 46.53 | 1.39340 | 39.89330 |
| O-F | 3.27 | 0.34 | 0.06 | 0 | 1 | 0 | 37.30 | 1.52390 | 33.35930 |
| O-I | 2.85 | 0.34 | 0.18 | 0 | 1 | 0 | 37.30 | 1.98940 | 46.11830 |
| O-CO2H | 2.95 | 0.37 | 0.45 | 0 | 1 | 0 | 74.60 | 1.08300 | 40.36060 |
| O-Cl | 2.94 | 0.37 | 0.23 | 0 | 1 | 0 | 37.30 | 2.03820 | 38.41130 |
| O-Br | 2.85 | 0.39 | 0.23 | 0 | 1 | 0 | 37.30 | 2.14730 | 41.10130 |
| O-NO2 | 2.17 | 0.71 | 0.78 | 0 | 1 | 0 | 86.96 | 1.70760 | 40.64280 |
| M-NH2 | 4.79 | -0.16 | -0.66 | 1 | 0 | 0 | 63.32 | 1.54820 | 37.80570 |
| M-CH3 | 4.24 | -0.07 | -0.17 | 1 | 0 | 0 | 37.30 | 1.69320 | 38.36730 |
| M-OH | 4.08 | 0.12 | -0.37 | 1 | 0 | 0 | 57.53 | 1.09040 | 35.42430 |
| M-OCH3 | 4.09 | 0.12 | -0.27 | 1 | 0 | 0 | 46.53 | 1.39340 | 39.89330 |
| M-F | 3.87 | 0.34 | 0.06 | 1 | 0 | 0 | 37.30 | 1.52390 | 33.35930 |
| M-I | 3.86 | 0.34 | 0.18 | 1 | 0 | 0 | 37.30 | 1.98940 | 46.11830 |
| M-CO2H | 3.54 | 0.37 | 0.45 | 1 | 0 | 0 | 74.60 | 1.08300 | 40.36060 |
| M-Cl | 3.83 | 0.37 | 0.23 | 1 | 0 | 0 | 37.30 | 2.03820 | 38.41130 |
| M-Br | 3.81 | 0.39 | 0.23 | 1 | 0 | 0 | 37.30 | 2.14730 | 41.10130 |
| M-CF3 | 3.79 | 0.43 | 0.54 | 1 | 0 | 0 | 37.30 | 2.40360 | 38.40330 |
| M-CN | 3.60 | 0.56 | 0.66 | 1 | 0 | 0 | 61.09 | 1.25648 | 38.11630 |
| M-NO2 | 3.45 | 0.71 | 0.78 | 1 | 0 | 0 | 86.96 | 1.70760 | 40.64280 |
| p-NH2 | 4.92 | -0.16 | -0.66 | 0 | 0 | 1 | 63.32 | 1.54820 | 37.80570 |
| p-CH3 | 4.34 | -0.07 | -0.17 | 0 | 0 | 1 | 37.30 | 1.69320 | 38.36730 |
| p-OH | 4.58 | 0.12 | -0.37 | 0 | 0 | 1 | 57.53 | 1.09040 | 35.42430 |
| p-OCH3 | 4.47 | 0.12 | -0.27 | 0 | 0 | 1 | 46.53 | 1.39340 | 39.89330 |
| p-F | 4.14 | 0.34 | 0.06 | 0 | 0 | 1 | 37.30 | 1.52390 | 33.35930 |
| p-I | 4.00 | 0.34 | 0.18 | 0 | 0 | 1 | 37.30 | 1.98900 | 46.11829 |
| p-CO2H | 3.51 | 0.37 | 0.45 | 0 | 0 | 1 | 74.60 | 1.08300 | 40.36060 |
| p-Cl | 3.99 | 0.37 | 0.23 | 0 | 0 | 1 | 37.30 | 2.03820 | 38.41130 |
| p-Br | 4.00 | 0.39 | 0.23 | 0 | 0 | 1 | 37.30 | 2.14730 | 41.10130 |
| p-CN | 3.55 | 0.56 | 0.66 | 0 | 0 | 1 | 61.09 | 1.25648 | 38.11630 |
| p-NO2 | 3.44 | 0.71 | 0.78 | 0 | 0 | 1 | 86.96 | 1.70760 | 40.64280 |

### Plots

boxplot(benzoate$pka, main="Spread of pKa Values", ylab="pKa values")



hist(benzoate$pka, main="Distribution of pKa Values", xlab="pKa Values")



### Outliers

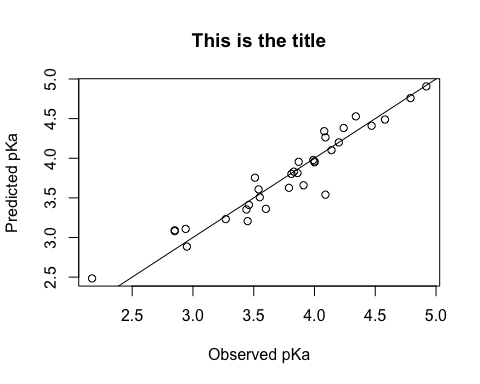
Based on the boxplot, are there any outliers in your pKa values? Which one? Look at your plot for model.2 – is this outlier a problem? Should we leave it in or remove it from the model. Explain your reasons.

## Model Building Results

### model.all results

Give the plot a title and copy it into this report.

model.all <- lm(pka ~ . , data=benzoate)  
plot(benzoate$pka, fitted(model.all), main="This is the title", xlab="Observed pKa", ylab="Predicted pKa")  
abline(a=0, b=1)



Use the summary() command to get the coefficients and Multiple R-squared. This also gives you Pr values for all the descriptors – you will need this for Part D.

summary(model.all)

##   
## Call:  
## lm(formula = pka ~ ., data = benzoate)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.31383 -0.14176 0.03044 0.05978 0.54947   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.3401515 0.3734497 11.622 2.42e-11 \*\*\*  
## sigma.m -0.7593948 0.4367495 -1.739 0.0949 .   
## sigma.p -0.6215509 0.2663965 -2.333 0.0283 \*   
## meta 0.0370935 0.2454376 0.151 0.8811   
## ortho -0.6856069 0.2528151 -2.712 0.0122 \*   
## para 0.1840806 0.2481625 0.742 0.4654   
## tpsa 0.0005219 0.0031450 0.166 0.8696   
## logp 0.0934268 0.1424129 0.656 0.5180   
## mr -0.0086522 0.0115723 -0.748 0.4619   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2051 on 24 degrees of freedom  
## Multiple R-squared: 0.9102, Adjusted R-squared: 0.8803   
## F-statistic: 30.42 on 8 and 24 DF, p-value: 9.617e-11

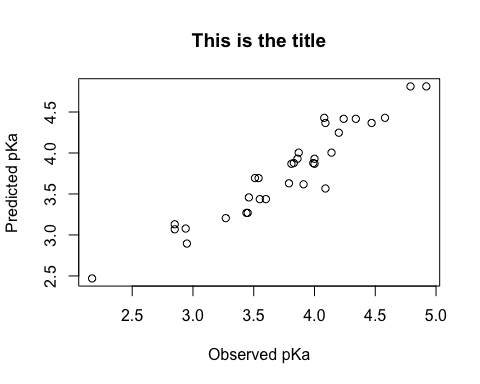
What is the equation for this model? e.g. pka = 31563146 + 3554first-descriptor – 2345second-descriptor + ……..

Discuss how good the fit is, including the Multiple R-squared. (a good fit will have R-squared close to 1.0, the closer the better)

### model.5 results

Give the plot a title and copy it into this report.

model.5 <- lm(pka ~ sigma.p + sigma.m + ortho + tpsa + logp, data=benzoate)  
plot(benzoate$pka, fitted(model.5), main="This is the title", xlab="Observed pKa", ylab="Predicted pKa")



Use the summary() command to get the coefficients and Multiple R-squared. This also gives you Pr values for all the descriptors – you will need this for Part D.

summary(model.5)

##   
## Call:  
## lm(formula = pka ~ sigma.p + sigma.m + ortho + tpsa + logp, data = benzoate)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.34892 -0.13829 0.00201 0.13166 0.52311   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.2197758 0.2864167 14.733 1.99e-14 \*\*\*  
## sigma.p -0.7106309 0.2400520 -2.960 0.00633 \*\*   
## sigma.m -0.5982669 0.3848878 -1.554 0.13174   
## ortho -0.7994722 0.0832319 -9.605 3.34e-10 \*\*\*  
## tpsa -0.0001302 0.0028483 -0.046 0.96388   
## logp 0.0233749 0.1045788 0.224 0.82482   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2065 on 27 degrees of freedom  
## Multiple R-squared: 0.8976, Adjusted R-squared: 0.8786   
## F-statistic: 47.32 on 5 and 27 DF, p-value: 1.601e-12

What is the equation for this model? e.g. pka = 31563146 + 3554first-descriptor – 2345second-descriptor + ……..

Discuss how good the fit is, including the Multiple R-squared. (a good fit will have R-squared close to 1.0, the closer the better)

### model.2 results

Give the plot a title and copy it into this report. Use the summary() command to get the coefficients and Multiple R-squared. This also gives you Pr values for all the descriptors – you will need this for Part D.

What is the equation for this model? e.g. pka = 31563146 + 3554first-descriptor – 2345second-descriptor + ……..

Discuss how good the fit is, including the Multiple R-squared. (a good fit will have R-squared close to 1.0, the closer the better)

### model.1 results

Give the plot a title and copy it into this report. Use the summary() command to get the coefficients and Multiple R-squared. This also gives you Pr values for all the descriptors – you will need this for Part D.

What is the equation for this model? e.g. pka = 31563146 + 3554first-descriptor – 2345second-descriptor + ……..

Discuss how good the fit is, including the Multiple R-squared. (a good fit will have R-squared close to 1.0, the closer the better)

## Discussion

3.1 In general, how does the ability of the model to predict the pKa compare as you increase the number of “descriptors” (model.1 → model.2 → model.5 → model.all)?  
Consider how the plots for the three models look and the Multiple R-squared values for the three models.

3.2 Statisticians will talk a parsimonious model as being preferred: use the simplest model that gives the best explanation. Given this, how does model.2 compare to the others? Do you think this is an improvement or not? Explain.

The original dataset had two outliers removed for you. You can’t just ignore data points that you don’t like – there needs to be a valid reason for leaving them out. Consider the chemistry of these molecules. Look closely at the structures of the molecules.

3.3 For *o*-Hydroxybenzoic acid the actual pKa is 2.98, but the model predicted a pKa of 3.77. What is the structure of *o*-hydroxybenzoic acid? Based on the structure, explain why the REAL pKa is lower than the model thinks it is – that is: it is a stronger acid than it “should” be according to the model. When this acid ionizes, the resulting carboxylate ion is more stable that you would think it should be – why? Note: the *meta* and *para* versions do not show this anomalous behavior.

3.4 For *o*-aminobenzoic acid the actual pKa is 4.98, but the model predicted a pKa of 4.09. What is the structure of *o*-Aminobenzoic acid?

Based on the structure, explain why the REAL pKa is higher than the model thinks it is. It is harder than it “should be” to remove a hydrogen ion from this molecule – explain why. Note: the meta and para versions do not show this anomalous behavior.

3.5 Since the model is describing the behavior of actual chemical compounds, it is important that the descriptors used in the model make “chemical sense” for the property they are meant to predict. Based on your knowledge of chemistry, why do you think the descriptors in model.2 are the best choices to explain pKa?

# References

Peter, Swathik Clarancia, Jaspreet Kaur Dhanjal, Vidhi Malik, Navaneethan Radhakrishnan, Mannu Jayakanthan, and Durai Sundar. 2019. “Quantitative Structure-Activity Relationship (Qsar): Modeling Approaches to Biological Applications.” In *Encyclopedia of Bioinformatics and Computational Biology*, edited by Shoba Ranganathan, Michael Gribskov, Kenta Nakai, and Christian Schönbach, 661–76. Oxford: Academic Press. <https://doi.org/https://doi.org/10.1016/B978-0-12-809633-8.20197-0>.