



Cluster-Based Regression Model for Predicting Aqueous Solubility of the Molecules

Priyanka Shit^(✉) and Haider Banka

Indian Institute of Technology (ISM) Dhanbad, Dhanbad, Jharkhand, India

priyanka15.ism@gmail.com

Abstract. Prediction of physicochemical properties is a crucial step in the drug discovery process. It is a combination of various tasks; one of the essential steps in that process is aqueous solubility prediction. Aqueous solubility ($\log S$) is a significant feature which is used to determine the drug-likeness of any compound. There are various machine learning, and statistical methods have been used to predict aqueous solubility in the literature. In this study, the aim is to propose a model which will improve the performance of the prediction model. In this study, data samples have clustered in different groups and built the regression model for each cluster. After that, the aqueous solubility value of each entity has predicted according to the cluster model. Combination of K-Means with various regression models has used for clustering and prediction purpose, respectively. Performance of the proposed model evaluated using Root Mean square error statistical measure. We have compared all the regression models with cluster-based model and got the best result with a cluster-based random forest model, which has RMSE value 0.6 and 0.61 for dataset 1 and dataset 2 respectively.

Keywords: Aqueous solubility · Drug-likeness · Clustering · Regression

1 Introduction

Aqueous solubility prediction is one of the challenging tasks among the drug discovery process and other applications. It can be used to determine the drug-likeness of any compound and also important for prediction of ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties [1, 2]. Human blood consists of 80% of water so, the compounds with low water solubility having low absorption rate. Therefore, prediction of aqueous solubility in an early phase of drug discovery and development process can help to reduce the time and cost factors. It also eliminates the molecules, which have very low aqueous solubility for reducing the risk of failure. Most of these are QSAR (Quantitative-Structure-Activity Relationship) methods where the activity of any molecule generate by mapping and encoding the structure of that molecule. Structures encode by undirected graph representation and mapping to specific length feature vectors. Chemical compounds can also be represented using SMILES notations, which is a standard representation and used to extract other information about that compound from chemical databases [3]. It is imperative to find a suitable regression or classification

model to learn the input training samples. There are various computational models have been applied to solve cheminformatics problems [4]. Literature reflects significant work in the area of multiple properties prediction. Literature reflects significant work in the area of properties prediction like Quantitative Structure-Property Relationships Models (QSPR) [5, 6], Use of Deep learning for a molecular graph representation of drug-like molecules [7], the role of solubility oral absorption prediction using a decision tree [8], Random forest model [9], theoretical models [10, 11], kinetic and intrinsic solubility prediction by potentiometric [12], multi-linear regression model, theoretical physics with machine learning models [13]. In this study, the aqueous solubility of molecules has predicted using six regression models. Here, the main aim to improve the performance of the predictive model. For this purpose, Data samples have clustered before applying to the predictive regression model. All molecules have represented in terms of molecular descriptors, which are the inputs for the predictive model. The descriptions of all methodologies are in Sect. 2. Aqueous solubility has predicted using the proposed model and also without clustering approach. Statistical performance measures have used to compare the simulation results of existing regression models with clustering approach.

2 Materials and Methods

2.1 Dataset

In this study, we have used two datasets; one of them is a small dataset of the only drug-like molecule, and another is a relatively large dataset with mixed data on drugs and non-drugs molecules. The details of each dataset have given in Table 1. We have collected SDF format, and molecular descriptors have extracted using the freely available PaDEL software.

Table 1. Dataset details

Dataset	Total samples	Training	Testing	Resource
Dataset 1	745	595	150	Ref [14]
Dataset 2	1708	1366	342	Ref [15]

2.2 Physicochemical Properties

All the molecules have collected in SDF format from the previous literature and manually from databases. Molecular descriptors can be estimated using various online server or using the software. In this study, we have used freely available PaDEL software for physicochemical properties calculation. Initially, we have taken 1185 features and after applying feature reduction methods the reduced feature set is 19. The reduced features, which have used for this study are XLogP, Hydrogen bond acceptor, Hydrogen bond donor, HybRatio, TopoPSA, AlogP, MlogP, Lipoaffinity index, Hydrogen atom count, Oxygen count, CrippenLogP, bpol, molecular weight, Mcgowan Volume, Kappa Indices, CrippenMR, FracC, Ring count, Rotatable bond count.

2.3 Feature Selection

Chemo-informatics deals with drug discovery process which involved a huge number of compounds with different types of chemical properties. So it is very important to reduce feature space to enhance the prediction performance. Feature reduction provides faster and cost-effective predictive algorithms and it also helps to understand the underlying process that generated the data which is a very important step in machine learning process. In order to reduce the no of irrelevant features in this study, relief and Pearson correlation methods have used for feature selection.

Relief: Relief is a feature selection algorithm which uses a statistical method. It can handle both discrete and continuous type features. It runs in low order polynomial time and independent to heuristic search. It takes linear time and effective when the instances are large in number. It computes weights and ranking of all the features for any input data samples using the target vector. This is effective for classification and also for the regression problem. The feature weight value is ranging from -1 to $+1$ [16, 17].

Pearson Correlation: Pearson Correlation is a basic feature reduction technique which is used to measure the linear dependency or correlation between two features [18]. The outcome of this method lies in the range of -1 to $+1$, both values are inclusive. Where $+1$ is for the positive correlation, -1 is for negative correlation and 0 is no correlation. The correlation is calculated by the following equation:

$$\gamma = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{n}) * (\sum Y^2 - \frac{(\sum Y)^2}{n})}} \quad (1)$$

Where X and Y are the features between which correlation is Calculated and n is the number of samples.

2.4 Clustering

Cluster analysis or clustering approach is based on an unsupervised learning approach in which the grouping of data samples takes place according to their similarity. Data samples in each group are more similar to the data sample of the same group as compare to other groups. Clustering is most important to visualize the data effectively. There are various types of clustering algorithms like connectivity models, centroid models, distributive models, density models etc. according to their application in the literature, K-Means algorithm is one of the popular centroid models clustering techniques.

K-Means Clustering Algorithm: It is based on an iterative process where the main aim of each iteration is to find local maxima. K-Means algorithm [18, 19] can be described by following steps:

- Step 1: Specify a desirable number of clusters.
- Step 2: Assign data points randomly in each cluster.
- Step 3: Compute centroids of each cluster.

- Step 4: According to closest cluster centroid reassign each data points.
- Step 5: Recompute cluster centroids.
- Step 6: Repeat steps 4 and 5 or terminates if no improvement possible and reach global optima.

2.5 Regression Analysis

A regression analysis widely used statistical method in which relationship among the variables estimated. It is used for prediction or forecasting so it also comes under as machine learning approaches. It helps to understand that what the effect on the dependent variable is when values of the independent variable are changed. In this method, problem is to find a variable or set of variables which are significant predictors of the required criterion variable. It may depend on a single independent variable or combination of a set of independent variables. There are various types of the regression model in the literature but in this study, we have used Gaussian Progress Regression, Multi-Linear Regression, Neural Net Regression, Ensemble Regression, Random Forest Regression, SVM Regression predictive models for comparative study [20–23]. All the given regression models have been used with and without cluster-based model for comparative study.

2.6 Statistical Performance Measures

In this study, to compare the result and performance Analysis of prediction model we have used Root Mean Square Error (RMSE) as Statistical performance matrices:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i^{Actual} - Y_i^{Predicted})^2} \quad (2)$$

3 Proposed Model

Aqueous solubility or other physiochemical properties have been predicted using various regression models in the previous studies. In this study, the main aim of the proposed model to increase the performance of the prediction model by clustering the data samples. The steps of the proposed model have been described in Fig. 1. So for this, in the first step, we have taken the solubility dataset which has been described in the previous section. All the data samples are represented in terms of molecular descriptors. Few molecular descriptors have been selected from the huge feature space for this study. For the selection of relevant features, there are various techniques like Filter, wrapper or subset selection. For this model, combination of relief and Pearson correlation method has been used as a feature selection method. Then data samples have been clustered according to their similarity. Clustering is important to visualize the data according to similarity and dissimilarity among them. Data samples which are more similar they belong to the same group or cluster. Regression analysis is a set of statistical processes

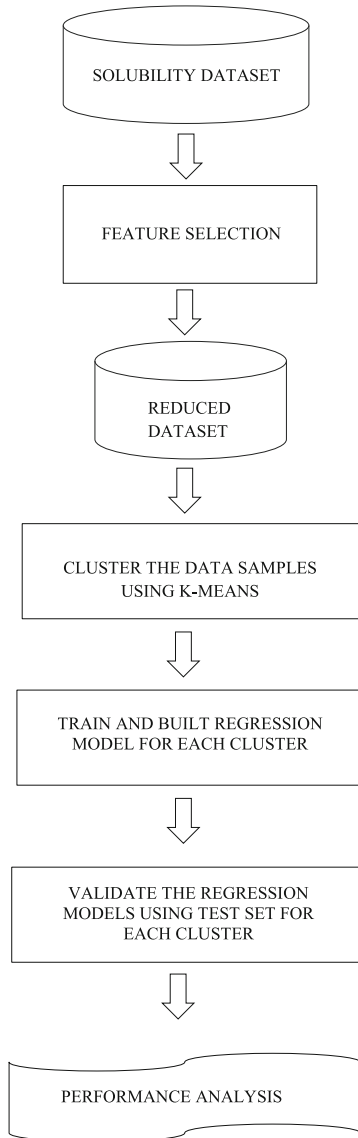


Fig. 1. Proposed cluster-based regression model

in which the value of the single variable is estimated in terms of other variables which is represented as the equation. Mapping all data in a single equation is very difficult for a large sample set. When the size of the data set is increased then mean absolute error is also increased. In regression analysis error is calculated in term of difference in actual and predicted value. So to reduce the error of predictive model data samples have been clustered before prediction in the proposed model. After clustering the data samples regression model is implemented for each cluster. The aqueous solubility of each new sample has been predicted after assigning to a specific cluster and the target

value is calculated according to the cluster equation. This step reduces the overall mean absolute error which improves the performance of the predictive model. It eliminates the limitation of a single regression model. In this model, K-Means have been used as a clustering technique and for prediction, various regression models have been used like Gaussian regression, multi-linear, neural net regression, ensemble regression, random forest regression, SVM regression. Using K-Means clustering data has divided into more than one group. After clustering for each group we have trained the model and finding different regression equation. For validation, the predicted value is calculated for each equation and finds the average and compares the result with actual log S value. In the last step for performance analysis, root mean square error (RMSE) has used as statistical performance matrices and compare the performance of other given regression model with the proposed cluster-based regression predictive model. The outcomes of this model have been described in the next section.

4 Result and Discussion

Aqueous solubility dataset 1 and dataset 2 collected from the resources was in SDF form, so features have extracted for this using PaDEL software. There is a huge set of physiochemical properties to represent any compound, so it is also a challenging task to select the properties which give more accurate values of other property. In this study, for feature selection, we have used relief feature selection and after that Pearson Correlation i.e. a combination of both methods have used. Relief is an effective model which is used to select best-ranked features and Pearson correlation has been used to eliminate redundancy of the feature set and also used to select most relevant features which are highly correlated with the target. So the combination of both techniques performs better with given predictive models. For prediction purpose, we have used 6 different regression models and compared the performance of each model with and without a proposed cluster-based model. All predictive models have trained and tested using MATLAB and WEKA software. The training and testing phase results for each dataset has represented in tabular form Table 2, Table 3 respectively.

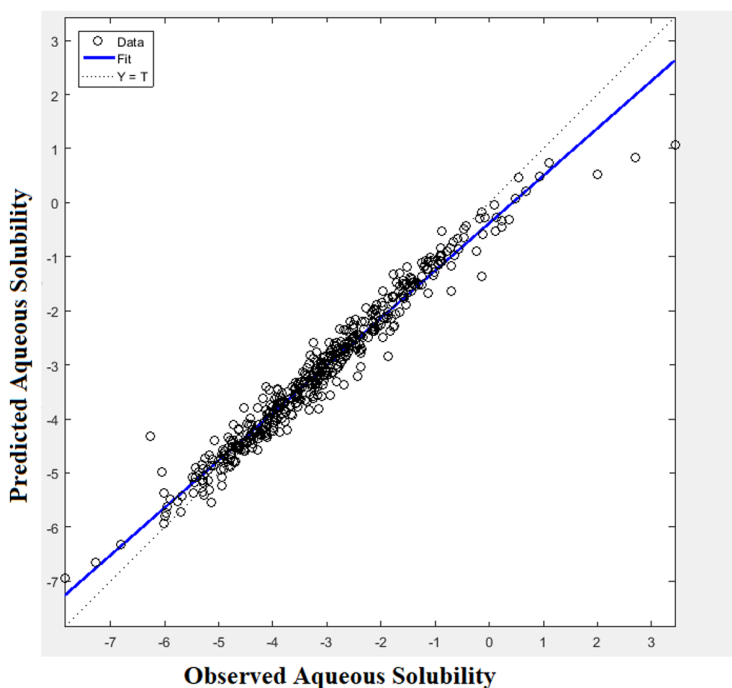
Table 2. Training and Test phase Phase results for Dataset 1 with and without a cluster-based model

Regression model	Without clustering		With clustering	
	Training RMSE	Testing RMSE	Training RMSE	Testing RMSE
Gaussian regression	0.70	0.81	0.65	0.75
Multi-linear regression	0.67	0.79	0.60	0.72
Neural net regression	0.66	0.80	0.60	0.78
Ensemble regression	0.65	0.81	0.65	0.80
Random forest	0.53	0.65	0.50	0.60
SVM regression	0.70	0.83	0.68	0.75

Table 3. Training and Test phase Phase result for Dataset 2 with and without a cluster-based model

Regression model	Without clustering		With clustering	
	Training RMSE	Testing RMSE	Training RMSE	Testing RMSE
Gaussian regression	0.77	0.79	0.63	0.65
Multi-linear regression	0.82	0.86	0.70	0.72
Neural net regression	0.84	0.88	0.53	0.79
Ensemble regression	0.55	0.77	0.45	0.66
Random forest	0.45	0.72	0.36	0.61
SVM regression	0.85	0.88	0.70	0.73

Table 2 is the result summary of dataset 1 training and test phase in which, we have got the best result with cluster-based random forest model. For dataset 1 best outcome RMSE, 0.60 is with using two cluster centers. Due to small data samples by increasing number of centers performance also decreases. Figure 2 and Fig. 3 is regression plot of training phase and test phase random forest model for dataset 1 respectively. Due to biased and small dataset of only drug-like molecules the result is not significantly improved using cluster based model.

**Fig. 2.** Regression plot of cluster-based Random Forest model dataset 1 training phase

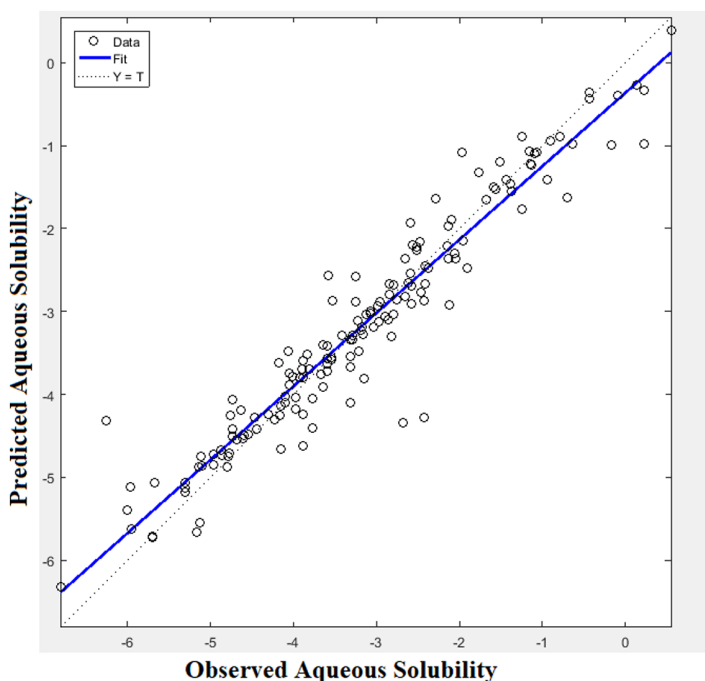


Fig. 3. Regression plot of cluster-based Random Forest model dataset 1 test phase

Therefore we have also taken another relatively large dataset of mixed molecules and compared the result. The original work has been done using dataset 2 and dataset 1 has taken only for comparative study for different sample size dataset. In Table 2, we can see that in the training phase error is less as compare to Table 3 but test result has not improved. Except random forest model using other model for dataset 1 performance is low. Due to small sample size if we increase the number of clusters predictive model is not perform effectively. For dataset 2 optimum number of clusters is 4, above this performance decreases relatively in this experiment. In Table 3, training phase error is very low for random forest model in both cases and also gives best result with this model. Figure 4 and Fig. 5 are the representation of regression plots for training and test phase of best model using actual and predicted solubility values.

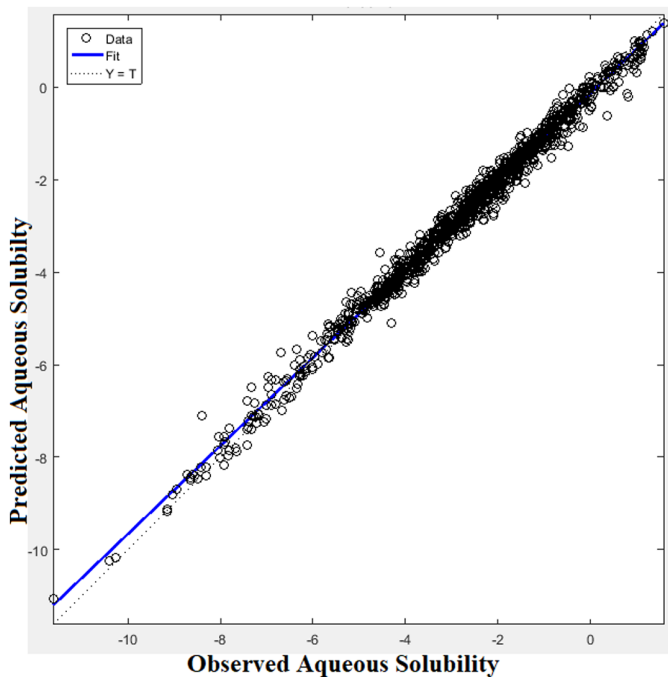


Fig. 4. Regression plot of cluster-based Random Forest model dataset 2 training phase

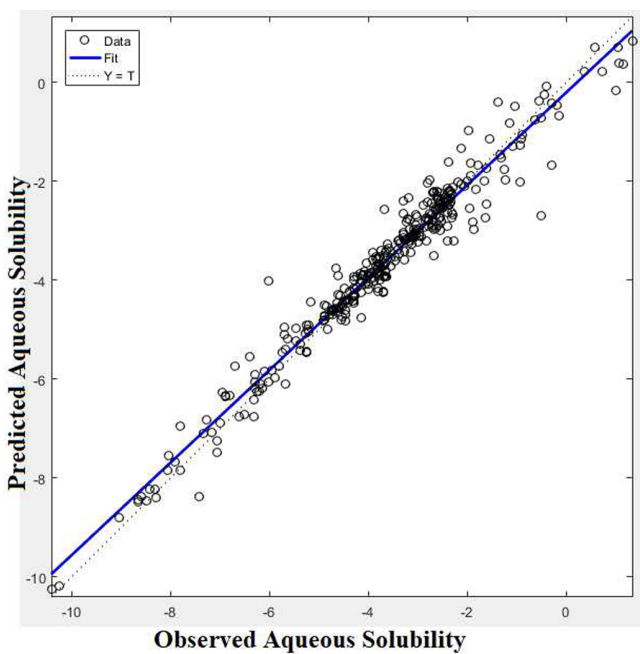


Fig. 5. Regression plot of cluster-based Random Forest model dataset 2 test phase

5 Conclusion

Predicting aqueous solubility is a critical task in the area of chemo-informatics. In this study, we have used two types of dataset one is a minimal sample size, and another is a relatively large sample. In both cases, with clustering or without clustering, we got the best result with a random forest regression model. We have got the lowest RMSE of 0.60 and 0.61 for dataset 1 and dataset 2 test phase respectively. For getting more improved unbiased results, a vast dataset should use. In this cluster-based model, we have got a good result with only using two clusters and four clusters for dataset 1 and dataset 2 respectively. For other datasets, the number of groups may vary.

References

1. Wang, J., Hou, T.: Recent advances on aqueous solubility prediction. *Comb. Chem. High Throughput Screen.* **14**(5), 328–338 (2011)
2. Murdande, S.B., et al.: Aqueous solubility of crystalline and amorphous drugs: challenges in measurement. *Pharm. Dev. Technol.* **16**(3), 187–200 (2011)
3. O’Boyle, N.M., et al.: Open babel: an open chemical toolbox. *J. Cheminform.* **3**(1), 33 (2011)
4. Habibi, N., et al.: A review of machine learning methods to predict the solubility of overexpressed recombinant proteins in *Escherichia coli*. *BMC Bioinform.* **15**(1), 134 (2014)
5. Hongmao, S.: *A Practical Guide to Rational Drug Design*. Woodhead Publishing, Cambridge (2015)
6. Schroeter, T.S., et al.: Estimating the domain of applicability for machine learning QSAR models: a study on aqueous solubility of drug discovery molecules. *J. Comput.-Aided Mol. Des.* **21**(9), 485–498 (2007)
7. Lusci, A., Pollastri, G., Baldi, P.: Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *J. Chem. Inf. Model.* **53**(7), 1563–1575 (2013)
8. Newby, D., Freitas, A.A., Ghafourian, T.: Decision trees to characterise the roles of permeability and solubility on the prediction of oral absorption. *Eur. J. Med. Chem.* **90**, 751–765 (2015)
9. Palmer, D.S., et al.: Random forest models to predict aqueous solubility. *J. Chem. Inf. Model.* **47**(1), 150–158 (2007)
10. Palmer, D.S., et al.: First-principles calculation of the intrinsic aqueous solubility of crystalline druglike molecules. *J. Chem. Theory Comput.* **8**(9), 3322–3337 (2012)
11. Palmer, D.S., et al.: Predicting intrinsic aqueous solubility by a thermodynamic cycle. *Mol. Pharm.* **5**(2), 266–279 (2008)
12. Narasimham, L., Barhate, V.D.: Kinetic and intrinsic solubility determination of some β -blockers and antidiabetics by potentiometry. *J. Pharm. Res.* **4**(2), 532–536 (2011)
13. McDonagh, J.L., Nath, N., De Ferrari, L., Van Mourik, T., Mitchell, J.B.O: Uniting cheminformatics and chemical theory to predict the intrinsic aqueous solubility of crystalline druglike molecules. *J. Chem. Inf. Model.* **54**(3), 844–856 (2014)
14. Kumar, R., et al.: Classification of oral bioavailability of drugs by machine learning approaches: a comparative study. *J. Comp. Interdisc. Sci.* **2**(9), 1–18 (2011)
15. Hou, T.J., et al.: ADME evaluation in drug discovery. 4. Prediction of aqueous solubility based on atom contribution approach. *J. Chem. Inf. Comput. Sci.* **44**(1), 266–275 (2004)
16. Wang, J., et al.: Development of reliable aqueous solubility models and their application in druglike analysis. *J. Chem. Inf. Model.* **47**(4), 1395–1404 (2007)

17. Urbanowicz, R.J., et al.: Benchmarking relief-based feature selection methods. arXiv preprint [arXiv:1711.08477](https://arxiv.org/abs/1711.08477) (2017)
18. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. *Comput. Electr. Eng.* **40**(1), 16–28 (2014)
19. Stanforth, R.W.: Extending K-means clustering for analysis of quantitative structure activity relationships (QSAR). Diss. University of London (2008)
20. Smits, G.F., Jordaan, E.M.: Improved SVM regression using mixtures of kernels. In: *Proceedings of the 2002 International Joint Conference on Neural Networks, 2002, IJCNN 2002*, vol. 3. IEEE (2002)
21. Kaytez, F., et al.: Forecasting electricity consumption: a comparison of regression analysis, neural networks and least squares support vector machines. *Int. J. Electr. Power Energy Syst.* **67**, 431–438 (2015)
22. Svetnik, V., et al.: Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **43**(6), 1947–1958 (2003)
23. Montgomery, D.C., Peck, E.A., Vining, G.G.: *Introduction to Linear Regression Analysis*, vol. 821. Wiley, Hoboken (2012)