UPPSALA
UNIVERSITET

U.U.D.M. Project Report 2015:10

Prediction of retention time

Oskar Gauffin

Examensarbete i matematik, 15 hp
Handledare: Tove Fall, Institutionen för medicinska vetenskaper
Ämnesgranskare och examinator: Silvelyn Zwanzig
Juni 2015

Department of Mathematics
Uppsala University

Contents

1 Introduction

Cardiovascular disease is a widespread health problem in large parts of the world. Recent studies of cardiovascular disease investigate the possibility to predict cardiovascular disease by molecules associated with metabolism, known as metabolites. An important problem in these studies is identification of metabolites, which is performed with the aid of several measurements. One of these is based on liquid chromatography and is called retention time. Retention time can be thought of as the passing time for a metabolite through a pipe. Since these measurements are expensive and time consuming, other ways of determining retention time for metabolites would be helpful.

2 Background and earlier studies

2.1 The classification problem

We describe the process of classifying metabolites. A chemist tries to annotate, i.e. classify, an unknown metabolite, for which a spectrogram, molecular weight (g/Mol) and retention time is known. The idea is to compare these measurements with a database containing previously annotated metabolites together with their properties e.g. weight and polarity measures. From this we try to classify the unknown metabolite as one of those in the database. More precise, we first condition on weight which is measured with high accuracy. However, it is not uncommon for several metabolites to share the same weight. The chemist then turns to spectrograms, which can be thought of as a partitioning of the molecule in smaller pieces, and tries to identify the metabolite using subject knowledge. However, sometimes the spectrograms are similar for several metabolites. Then we would want to use our knowledge of retention time for the unknown metabolite, but retention time is not registered in the database. Therefore we need to accurately predict retention time, from other characteristics in the database.

2.2 Earlier studies

Hagiwara et al (2010) reported a model for classifying metabolites using predicted retention time, using a training set of 150 metabolites with known retention time (rt). For higher ($rt > 90$ s.) retention times they observe a linear relation between retention time and two variables, log of the partition coefficient ($\log p$), and a solvent accessible surface area, and use multiple linear regression for these metabolites. For metabolites with shorter retention times they instead use support vector regression.

2.3 The data

Retention time of 611 classified metabolites were measured in a Waters Acquity UPLC BEH C8 column using a solution with 95 % water and 5 % methanol, which is then continuously shifted towards a higher methanol concentration until opposite relation between water and methanol is achieved. Of these measured retention times, 3 where discarded after consulting a chemist, since their retention times were considered too high to be accurately measured.

A problem with the data collection is that the selected 608 metabolites were not randomized over the set of all possible metabolites. Instead they were selected from several different criteria in earlier studies, hence some of the metabolites might not be well represented by our data.

Except for the retention time, all data used for the predictions is found in the Human Metabolome Database[1] (HMDB). HMDB contains a variety of measures and predictions for 41514 metabolites found in humans. However many metabolites lack measurements, i.e. experimental measurements have not yet been made for all metabolites. Therefore we decided to use predicted measurements, which are available for most metabolites. These predicted measurements describes the polarity and mass of the metabolite. Several of the measurements are highly correlated, e.g. the log of the partition coefficient is predicted twice with two different methods, which naturally makes these two predictions correlated. We briefly describe each selected explanatory variable, and describe the variable selection below in the method section.

- *log of the partition coefficient (*$\log p$*)* measures the solubility of a metabolite. Predictions of the $\log p$ are based on experiments dissolving the metabolite in octanol and water. The quotient of the concentration of the metabolite in octanol divided by the concentration of the metabolite in water, is the partition coefficient. In this study we choose to work with predictions from both the ALOGPS method, as described in Tetko et al (2005) and the chemaxon[2] method.

- *polarizability* measures how the metabolite responds to electrical fields.

[1] See http://www.hmdb.ca/

[2] See https://www.chemaxon.com/ for details

- *pka of strongest acid* is the negative of the log of the acid dissociation constant K_a, which is the equilibrium point where the acid HA, H_3O^+ and A^- does not change in concentration over time. This measures the acidity of the metabolite, where higher values indicate weaker acids.
- *pka of strongest base* is the corresponding basic measure, indicating how basic the metabolite is.
- *log of solubility (*$\log s$*)* measures the water solubility of the metabolite.
- *class* is the class to which the metabolite belongs.

Remark. *It is known that these predicted values are not always accurate. The* $\log p$*-predictions from ALOGPS for instance has a reported*[3] *root mean squared error of 0.35, and typically takes values within the range of -10 to +10. This means there are non-trivial errors in our variables.*

3 Theory

3.1 Ordinary Least Squares

Ordinary Least Squares Regression (OLS) is the most well known type of regression, assuming we have predictor variables $X = (X_1, X_2, \ldots, X_p)$ and a response variable y, for which the following relation holds

$$y = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \textit{where}\, r(X) = p.$$

For this relation we require the Gauss Markov Conditions

$$\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}, \tag{1}$$

$$\mathrm{E}(\epsilon_i^2) = \sigma^2 < \infty, \tag{2}$$

$$\mathrm{E}(\epsilon_i \epsilon_j) = 0 \text{ when } i \neq j \tag{3}$$

for all $i,\ j = 1, ..., n$.

Minimizing $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 X_{1i} - \ldots - \beta_p X_{pi})^2$ for estimates of $\boldsymbol{\beta}$ gives the OLS-estimate $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T y$. From this we define the hat matrix $H := X(X^T X)^{-1} X^T$ and $M := (I - H)$. We will now make some preparations to show that under the GM-conditions above, $\hat{\boldsymbol{\beta}}$ is in some sense the best possible estimate of $\boldsymbol{\beta}$.

[3] "The logP prediction accuracy is root mean squared error rms=0.35" from http://www.vcclab.org/lab/alogps/

We define the residuals of a fit as $\boldsymbol{r} := y - \hat{y} = (y_1 - \hat{y}_1, \ldots, y_n - \hat{y}_n)$. From the following connection between $\boldsymbol{\epsilon}$ and $\boldsymbol{r}$,

$$\boldsymbol{r} = y - X(X^T X)^{-1} X^T y = (I - H)y = My = MX\,\boldsymbol{\beta} + M\boldsymbol{\epsilon} = M\boldsymbol{\epsilon},$$

we show that the covariance of the residuals are closely connected to $\mathrm{Var}(\epsilon_i) = \sigma^2$, i.e.

$$\mathrm{Cov}(\boldsymbol{r}) = \mathrm{Cov}(M\epsilon) = M\sigma^2 M = \sigma^2 M. \tag{4}$$

In the last equality we used that M is idempotent. This implies that M is a covariance matrix, hence positive semidefinite.

We now give the main argument for using $\hat{\boldsymbol{\beta}}$ estimates, as presented in Sen & Srivastava (1990). Let L be an arbitrary matrix. We call functions $L\,\boldsymbol{\beta}$ of $\boldsymbol{\beta}$, for which linear unbiased estimates exists, estimable. Surely, we could estimate $L\,\boldsymbol{\beta}$ in a nonlinear or biased way as well. But when $L\,\boldsymbol{\beta}$ is estimable the following theorem tells us that among the class of unbiased, linear estimates $L\,\hat{\boldsymbol{\beta}}$ has the smallest variance possible.

Theorem 1. *(Gauss Markov Theorem)*
Let $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T y$, $y = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $L\,\boldsymbol{\beta}$ be an estimable function of $\boldsymbol{\beta}$, $r(\boldsymbol{X}) = p$ and assume that the GM-conditions hold. Then $L\,\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimate of $L\,\boldsymbol{\beta}$, where best is understood as $\mathrm{Cov}(Cy) - \mathrm{Cov}(L\,\hat{\boldsymbol{\beta}})$ being positive semi-definite, where Cy is any another linear unbiased estimate of $L\,\boldsymbol{\beta}$.

Proof. Since Cy is unbiased, $\mathrm{E}(Cy) = L\,\boldsymbol{\beta}$, and $\mathrm{E}(Cy) = CX\,\boldsymbol{\beta}$ for all $\boldsymbol{\beta}$, hence $L = CX$. Now $\mathrm{Cov}(Cy) = \sigma^2 CC^T$ and $\mathrm{Cov}(L\,\hat{\boldsymbol{\beta}}) = \sigma^2 L(X^T X)^{-1} L^T$. From this

$$\mathrm{Cov}(Cy) - \mathrm{Cov}(L\,\hat{\boldsymbol{\beta}}) = \sigma^2 C(I - X(X^T X)^{-1} X^T)C^T, \tag{5}$$

where we recognize $(I - X(X^T X)^{-1} X^T)$ as the covariance matrix M of the residuals, therefore (5) is semidefinite positive. □

3.2 Examining the conditions

The Gauss-Markov conditions must of course be assessed. As we have seen before, the residuals are closely connected to the unobservable ϵ, which suggests examining the residuals for patterns. If the first GM-condition is violated, the positive and the negative residuals will be unevenly distributed. If the second GM-condition is violated, the variance of the data is not constant, a phenomenon known as heteroscedasticity. This might affect estimates of the variance, which for instance causes problems in some model diagnostics. Violating the third GM-condition could result in patterns in the residuals, e.g. suggesting a quadratic term is appropriate.

From (4) it follows that $\operatorname{Var} r_i = \sigma^2(1 - h_{ii})$, where, h_{ii}, the so called leverage, denotes the diagonal elements of H. Since the variance of the residuals depends on both leverage h_{ii} and σ^2, we might want to standardize the residuals before examining them, following this definition from Sen & Srivastava (1990).

Definition 1. *Studentized residuals is defined as*

$$r_i^* = \frac{r_i}{s_{(i)}\sqrt{1 - h_{ii}}},$$

where $s_{(i)}$ is the standard deviation of the data without the i:th point.

Examining the residuals plotted against predicted response and predictors are common ways to assess these conditions.

Another problem that might turn up in a residual plot is the presence of outliers, that is points which differ from the majority of the data in some sense. Such points might bias the fit, especially if they are in a leverage position, i.e. it has a larger h_{ii}-value than other points. However, since the hat matrix does not take the response into account, and since outliers could mask other outliers, simply looking at leverage is not sufficient to decide whether a point is biasing the fit. Neither does examining residuals suffice to pinpoint outliers. An outlier could pull the fit towards it, causing both a biased fit, and a small residual for itself. Therefore we will present other tools in the next section.

3.3 Robust methods

Classical statistical theory, including the OLS-estimate, has been criticized for putting to much emphasis on assumptions which real data usually do not fulfil, for instance in Huber (1981). These critics suggest we use methods which performs well even for moderate deviations from assumptions, and call these methods robust. The downside of being robust is lower Gaussian efficiency compared to the classical counterparts, i.e. efficiency under normally distributed data. For completeness we give the following definition of efficiency from Liero & Zwanzig (2011).

Definition 2. *Efficiency of an unbiased estimator T of a parameter θ is defined as*

$$e(T,\theta) = \frac{1/I(\theta)}{\mathrm{Var}(T)},$$

where $I(\theta)$ is the Fisher information.

From a more practical point of view, we could settle for better predictions for a majority of the data, on the behalf of worse predictions for some outliers.

It is well known that OLS-estimates are sensitive to outliers since it minimizes $\sum_i^n r_i^2$, i.e. gives outliers quadratic influence. There are many options, e.g. Maronna et al (2006) present the least absolute deviation (LAD), where instead $\sum_i^n |r_i|$ is minimized. However such an estimate is in a certain sense not more robust, since the contamination of a single outlier still has unbounded influence on the fit. This idea is formalized in the next two definitions, from Rousseeuw & Leroy (1987).

Definition 3. *We define the maximum bias of the regression estimator T applied to a sample X and contaminated samples X^* where c points are contaminated, i.e. replaced with arbitrary values, as*

$$B(c;T,X) = \underset{X^*}{sup}||T(X^*) - T(X)||,$$

where we use euclidean norm and subscript X^* denotes that the supremum is taken over all possible contaminated sets.

Definition 4. *We define the breakdown point as the smallest contamination c for which $B(c;T,X)$ can become arbitrarily large, i.e.*

$$\epsilon_c^*(T,X) = min\{0 \leq \frac{c}{n} \leq 1 : B(c;T,X) = \infty\}.$$

Remark. *Note that all breakdown points fulfil $0 \leq \epsilon_c^*(T,X) \leq 0.5$, when $n \to \infty$, which is referred to as asymptotic breakdown point.*

In this regard, OLS and LAD are equally poor estimators, since their asymptotic breakdown point is $1/n \to 0$ as $n \to \infty$. For a higher breakdown point, Rouesseew (1984) suggests interchanging the summation sign in OLS and LAD with the sample median, which results in

Definition 5. *The Least Median Squares (LMS) is defined as*

$$\underset{\hat{\boldsymbol{\beta}}}{argmin}\ med\ r_i^2.$$

Remark. *The LMS-estimate is geometrically a hyperplane parallel to the thinnest (i.e. thinnest in vertical direction) hyperstrip containing at least* $[n/2] + 1$ *of the points, where square brackets denotes the closest integer-function.*

The LMS estimate has several nice theoretical properties, e.g. the highest possible asymptotic breakdown point 0.5. However, the exact estimates for LMS for moderate to large data sets in high dimensions requires heavy computations. Instead of searching through all subsets of the data, a compromise is usually made, and an approximate solution is found. These approximations are criticised in Hawkins and Oliver (2002), for not producing the intended high breakdown point estimates, especially so when the number of initial samples is small.

We will now motivate and describe an algorithm described in Rousseeuw & Hubert (1996), which is referred to for details in the R package MASS lqs-function, which we use for LMS-regression. The algorithm assumes that the data is in general position, that is that no subset of p x_i lie on a $p-1$ dimensional hyperplane, where p equals the number of predictors plus 1 for the intercept. For most continuous data, general position is not an unreasonable assumption.

The algorithm tries to generate at least one contamination free sample of p points, with a high probability P_{clean}, by selecting many initial samples. Rousseeuw and Leroys (1987) suggests using a $P_{clean} = 0.99$. Following their reasoning, P_{clean} is hypergeometrically distributed, but if we only consider cases where n/p is large, we can approximate this with binomial distribution. Hence P_{clean} depends on the proportion of contamination ε, sample size p and number of samples m as

$$P_{clean} = 1 - (1 - (1 - \varepsilon)^p)^m.$$

For our purposes, setting $p = 6$, $\varepsilon = 0.5$, $m = 10^4$ yields $1 - P_{clean} < 10^{-68}$ which should suffice. We now give the algorithm.

1. Randomly select p observations.

2. We fit a hyperplane through the subset. Since the data is assumed to be in general position, this is always possible. If the sampled points are not in general position, we take another sample.

3. Adjust the intercept. This is done by selecting the midpoint of the shortest interval in y which contains $h = [(n/2] + 1$ points. If several

intervals have the same length, the median of these midpoints is taken as intercept.

4. Evaluate each fit based on the LMS minimization criteria, i.e. $medr_i^2$.
5. Iterate m times. Select the fit with smallest minimization criteria from step 4.

However it can be shown that LMS has a relatively low Gaussian efficiency of $n^{1/3}$, which can be compared to the well known Gaussian efficiency of OLS $\sqrt{n}$. Therefore Rousseeuw & Leroy (1987) propose using LMS as a preliminary weighting function. First a LMS-fit is made, from which the residuals r_{LMS} are used for a scale estimate

$$\hat{s} = C(1 + 5/(n - p))\sqrt{med(r_{LMS}^2)},$$

where $C = 1/\Phi^{-1}(0.75) \approx 1.4826$, since C makes $med|r|$ a consistent estimator of σ when $r \sim N(0, \sigma^2)$. $(1 + 5/(n - p))$ is a finite sample correction, where n is the number of observations, and p as before is the number of predictors + 1 for the intercept. We now introduce weight w_i as

$$w_i = \begin{cases} 1 & \text{if } |r_{LMS}/\hat{s}| \leq 2.5 \\ 0 & \text{otherwise,} \end{cases}$$

where 2.5 is chosen as a threshold simply since few residuals are larger than 2.5σ under the assumption of normal distribution. Next we perform a reweighted least squares, which minimizes $\sum_i^n w_i(y_i - \beta_0 - \beta_1 X_{1i} - \ldots - \beta_p X_{pi})^2$. Since the weights are either 1 or 0, this method can be thought of as an OLS-estimate where points with large residuals have been removed. A historical reason for using LMS was relatively cheap computational complexity, until Rousseeuw & van Driessen (2006) suggested an algorithm which reduces the computational cost of the least trimmed squares estimate, proposed in Rousseeuw & Leroy (1987).

Definition 6. *The Least Trimmed Squares (LTS) is defined as*

$$\underset{\hat{\boldsymbol{\beta}}}{argmin} \sum_{i=1}^{h} r_{(i)}^2,$$

where $0 < h \leq n$, and $r_{(i)}$ is the ith smallest residual. This means we find the OLS-fit to the h smallest squared residuals. We denote LTSXX as the least trimmed squares where we trim of 100-XX % of the residuals, e.g. LTS95 means that 5 % of the residuals were trimmed. It can be shown that LTS has a Gaussian efficiency of $\sqrt{n}$, and by trimming $h = [(n+p+1)/2]$ it attains 0.5 as asymptotic breakdown point. That being said about the Gaussian efficiency, Croux & Rousseeuw (1994) claim that the Gaussian efficiency of the LTS is still below 8 %.

We now present the relevant parts of the fast-LTS-algorithm, as presented in Rousseeuw & van Driessen (2006), implemented in the R package robustbase as "ltsReg". As before we assume data to be in general position, and specify the amount of trimming, i.e. pick an h such that $[n+p+1]/2 \leq h \leq n$.

1. Randomly select p points and fit a hyperplane to these points. If the rank of the subset is less than p, randomly select new points until rank p is achieved.

2. First concentration step. Calculate the residuals for each point from the fitted hyperplane, and sort them in increasing order. Fit a new OLS- hyperplane to the points with the h smallest residuals.

3. Second concentration step. Fit a new OLS-hyperplane to the h points with smallest residuals with respect to the OLS-hyperplane in step 2.

4. The intercept $\hat{\beta}_p$ in $\hat{\boldsymbol{\beta}}$ is adjusted into $\hat{\beta}_p^*$ before each LTS criterion check, and is found from $t_i = y_i - x_{i,1}\hat{\beta}_1 - \ldots - x_{i,p-1}\hat{\beta}_{p-1}$ by

$$\hat{\beta}_p^* = \underset{\mu}{argmin}(\sum_{i=1}^{h}((t_i - \mu)^2)_{i:n}),$$

   where the i:n subscript denotes that the $(t_i - \mu)^2$ are ordered in increasing size.

5. Iterate the steps above m times. Evaluate each fit with the LTS criterion, i.e. $\sum_i^h r_{(i)}^2$, and proceed with the 10 best.

6. Carry out concentration steps on each of the 10 fits until convergence, evaluate with LTS criterion and report the best fit.

We conclude this section by noting that there are many other robust regression techniques than those presented here, and it is not clear which one you should select. The following quote from Tukey, found in Huber (2002) p. 1645, sheds some light upon this. "[W]hich robust/resistant methods you use is not important - what is important is that you use some. It is perfectly proper to use both classical and robust/resistant methods routinely, and only worry when they differ enough to matter."

3.4 Errors in variables

We now address the presence of errors in our variables. Errors in the predictors give rise to so called attenuation bias, i.e. the coefficients are shrunken towards zero. For simplicity we consider a proof for simple linear regression, from Chen et al (2007), where we as usual assume the following model

$$y = \beta_0 + \beta_1 \tilde{x} + \varepsilon,$$

where $\mathrm{E}(\tilde{x}\varepsilon) = 0$. However, $\tilde{x}$ is observed with some measurement error, i.e. $x = \tilde{x} + e$ where $e \sim (0, \sigma_e^2)$ for which it holds that $\mathrm{Cov}(e, \varepsilon) = 0$ and $\mathrm{Cov}(\tilde{x}, e) = 0$. We now regress the error-prone version of $\tilde{x}$ on y, as

$$\hat{\beta}_1 = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n (x_i - \bar{x})^2}$$

Since

$$\frac{\mathrm{Cov}(y, x)}{\mathrm{Var}(x)} = \frac{\mathrm{Cov}(\beta_0 + \beta_1 \tilde{x} + \varepsilon - \beta e, x)}{\mathrm{Var}(\tilde{x} + e)} = \beta_1 (1 - \frac{\mathrm{Cov}(e, x)}{\mathrm{Var}(\tilde{x}) + \mathrm{Var}(e)}),$$

and since $Cov(e, x) = \mathrm{E}[e^2] = \mathrm{Var}(e)$, it follows that when sample size grows

$$\hat{\beta}_1 \underset{p}{\rightarrow} \beta_1 \frac{\sigma_{\tilde{x}}^2}{\sigma_{\tilde{x}}^2 + \sigma_e^2}.$$

This implies that coefficients derived under these conditions will be subject to shrinkage. However since the purpose of this study is to make predictions, we choose to believe Carroll et al (2006) p.19 who claim that "it rarely makes any sense to worry about measurement errors" when predictions are made on error-prone version of the predictors.

3.5 Multicollinearity

In order for $\hat{\boldsymbol{\beta}}$ to be unique, $(X^T X)^{-1}$ must exist, hence X must be linearly independent. However if X is close to linearly dependent an issue known as multicollinearity might occur. Chatterjee and Price (1977) p.155 presents examples which indicate that multicollinearity "can seriously limit the use of regression analysis for inference and forecasting". Sen and Srivastava (1990) gives a more detailed motivation, using the following lemma.

Lemma 1. *(Extended Cauchy-Schwarz Inequality)*
Let b be a vector, and B a positive definite matrix. Then

$$(b^T b)^2 \leq (b^T B b)(b^T B^{-1} b).$$

Proof. The proof can be found in Johnson and Wichern, (2007) p. 79. □

We characterize multicollinearity, i.e. $(X^T X)$ near singular, with the existence of a unit vector $\boldsymbol{c}$ for which $\boldsymbol{c}^{\boldsymbol{T}} X^T X \boldsymbol{c} = \epsilon$ where ϵ is small. Using the lemma, and assuming that $X^T X$ in $\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 (X^T X)^{-1}$ is not only semi-definite positive but definite positive, we get

$$1 = (c^T c)^2 \leq (c^T (X^T X) c)(c^T (X^T X)^{-1} c) = \epsilon (c^T (X^T X)^{-1} c) \Leftrightarrow$$

$$\mathrm{Var}(c^T \hat{\boldsymbol{\beta}}) = \sigma^2 c^T (X^T X)^{-1} c \geq \sigma^2 / \epsilon.$$

This means that the variance of $\hat{\boldsymbol{\beta}}$ is magnified in the presence of a small ϵ.

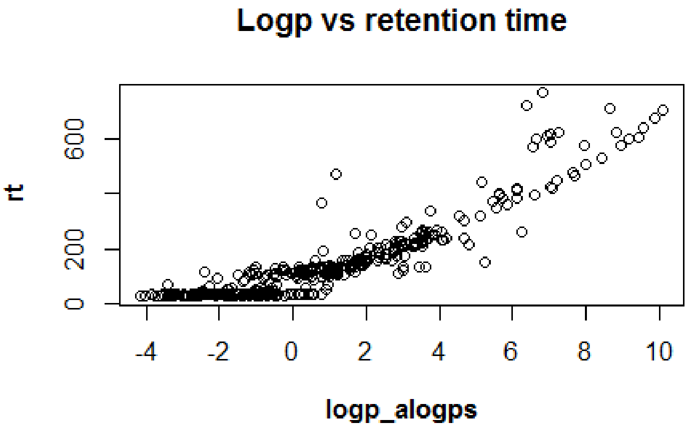
Several ways of detecting multicollinearity have been suggested. Here we settle for the so called Variance Inflation Factor (VIF), as presented in Sen & Srivastava (1990). First, we define R_j^2 as the R^2 of the linear regression of all predictor but the jth, on the jth predictor used as response. From this we define the VIF of the jth predictor as

$$VIF_j = \frac{1}{1 - R_j^2} \quad j \in \{1, \ldots, p\}.$$

A rule of thumb is that VIF greater than 5 indicates a multicollinearity problem, according to Rogerson (2001).

4 Variable and model selection

We begin by randomly splitting the data into a training set of 426 metabolites and a validation set with 182 metabolites, i.e. 30 % of the observations, and only use our validation set in the results-section. From plotting each predictor against the response, we observe the following pattern for the predictor $\log p$ from ALOGPS.

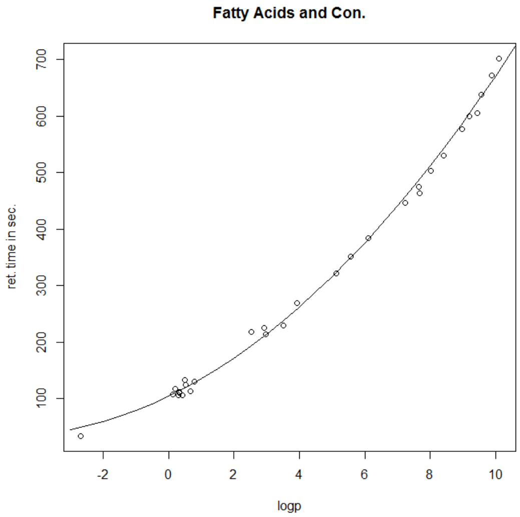


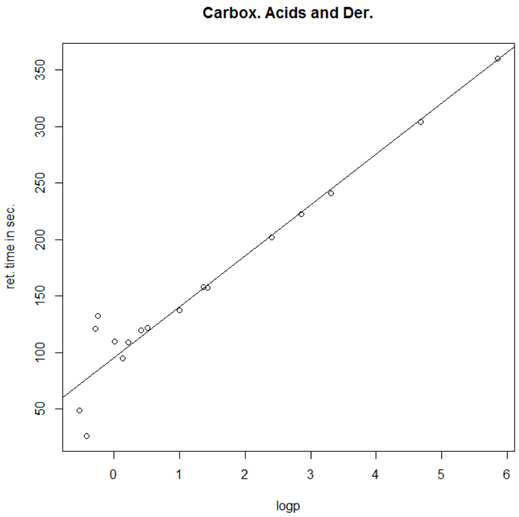
For most of the metabolites with rt-values below 45 seconds, the relationship between rt and $\log p$ has a small, if any, slope. For values above this threshold there is more curvature and higher variation. Even though higher $\log p$-values seems to push the retention time of a metabolite into the upper of these two clusters, the overlap between $-1 < \log p < 2$ suggests we could perhaps do better using more predictors.

First however, we deal with certain well behaved metabolites. Plotting metabolites belonging to the same metabolite class, we find that retention time in two classes are well explained by using $\log p$_alogps as single predictor. We fit a quadratic model to the first class, named ”Fatty Acids and Conjugates” (FA),

$$y = \beta_0 + \beta_1 \log p + \beta_2 \log p^2,$$

which gives us an $R^2 = 0.99$. The second class ”Carboxylic Acids and Derivatives” (CA) has a slightly lower $R^2 = 0.94$, probably caused by the overlap issue described above. However, with only one predictor it seems reasonable to conclude that the fit is not distorted in any severe way, and we settle for the simple OLS-fit here.



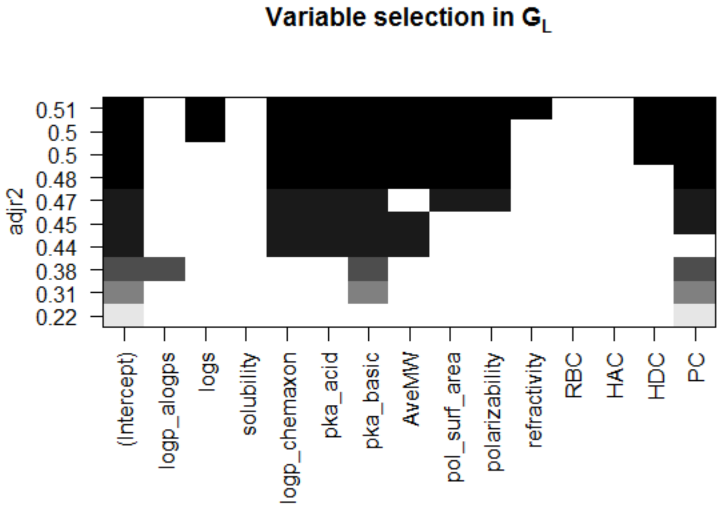


Some care is required when using these models for lower values of $\log p$. For the CA-class the increased variance is probably due to the overlap issue, and it would be bold to claim that the FA-class does not suffer from the same problem based on a single low rt-point. We try to counter this by using these models for the appropriate classes when $\log p$_alogps > 0.

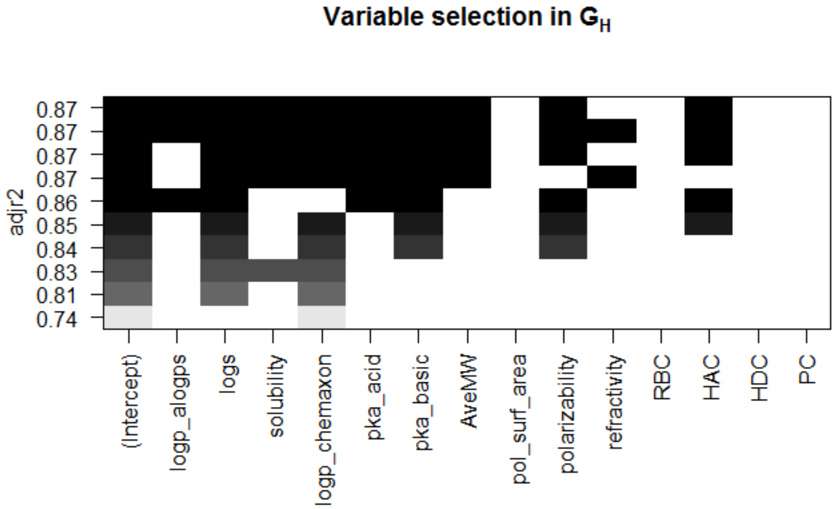
From here we remove points where $0 < \log p$ from the FA- and CA-class, and address the rest of the data. Similar to Hagiwara et al (2010), we suggest taking advantage of the classification context, i.e. knowledge of the retention time of the metabolite we are trying to classify. We split the training data into two groups G_H and G_L depending on whether $rt > T = 45$ s. Then for each unknown metabolite we condition on its retention time, and predict retention time using a low-rt-model $\hat{\boldsymbol{\beta}}_{low}$ if the metabolite is in G_L, and using a high-rt-model $\hat{\boldsymbol{\beta}}_{high}$ if the metabolite is in G_H.

4.1 Variable selection

We now turn to the problem of picking predictors for each model, using exhaustive variable selection. Preferably our model selection should have been as robust as our regression tools, however there was not time. Below is a plot displaying models with highest adjusted R^2 for a each $p \in 1, \ldots, 10$ predictors in the higher rt-group. We include variables as long as the adjusted R^2 increase more than trivially, but do not include a variable if it makes the VIF take values over 5. After an initial run, we iterate the exhaustive search including quadratic terms for previously selected variables which plotted against retention time display a quadratic relationship. This results in



This suggests using $\log p$_alogps, $\log p$_chemaxon, pka of strongest acid, pka of strongest basic and polarizability as predictors, which gives $R^2_{adj} = 0.43$ and $VIF = 4.2$. For the higher group the R^2_{adj} is larger as displayed in the following graph



This suggests that for G_H we use $\log p$_alogps, $(\log p\text{\_chemaxon})^2$ and $(\log s)^2$ as predictors. This gives an $R^2_{adj} = 0.90$ and a $VIF = 4.74$.

4.2 Model evaluation

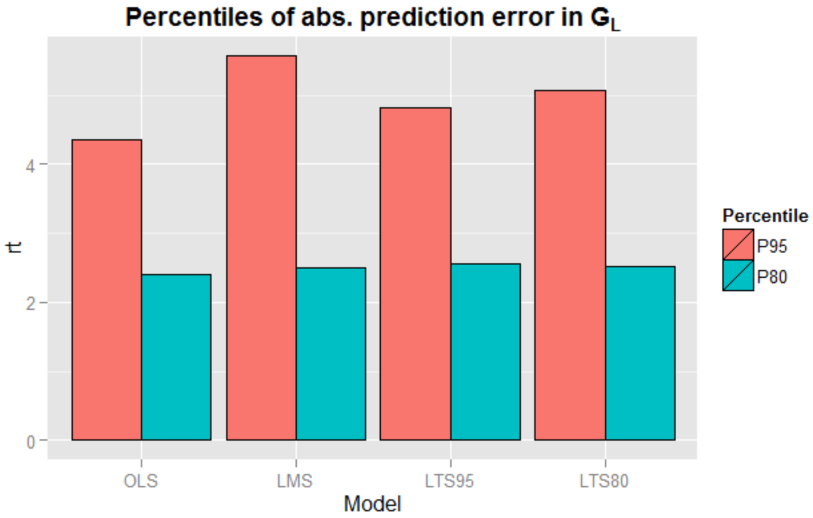
It is not trivial how to validate the predicted retention times since we do not know where precision is needed. The chemist might classify certain metabolites well using only weight and spectrograms, and require precise predictions of retention time in other cases. However from the classification context, an estimate of the upper bound of error for most predictions is of interest. If a normality assumption could be supported, we could use a prediction interval for the OLS-model. However, diagnostic plots and attempts with transformations have not supported such an assumption, hence we must try something else. Therefore we choose to proceed with the 80:th and 95:th sample percentile of the absolute prediction error, i.e.

$$P_{XX}(|y_i - X_i\,\hat{\boldsymbol{\beta}}_{H/L}\,|),$$

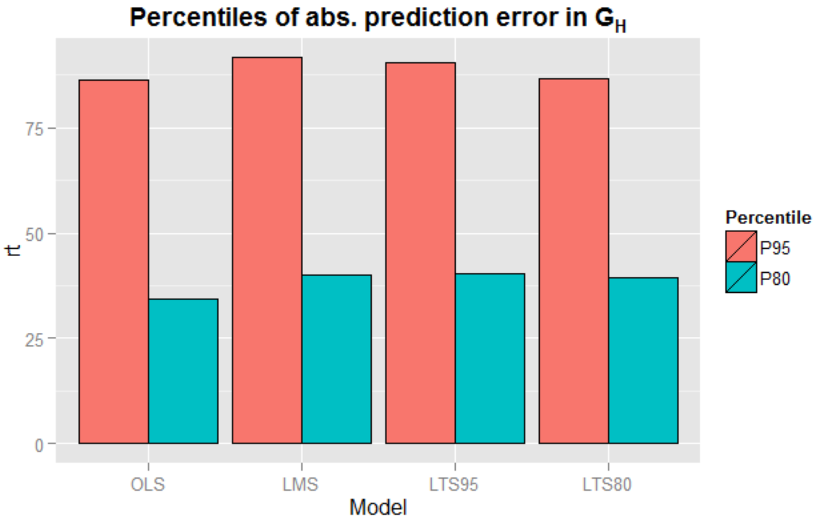
where $XX \in \{80, 95\}$ $X_i\,\hat{\boldsymbol{\beta}}_{H/L}$ denotes that we pick predictions made with $\hat{\boldsymbol{\beta}}_{low}$ for metabolites in G_L and $\hat{\boldsymbol{\beta}}_{high}$ for G_H. We denote these sample percentiles P_{80} and P_{95}.

4.3 Model selection

Using these variables we build four models for each of G_L and G_H: OLS, OLS weighted with LMS (which we will denote "LMS" in graphs), LTS95 and LTS80. Since we want to avoid overfitting our data, by selecting a model based on performance on the validation set, we instead validate these models on our training data using a five fold cross validation with 10^4 initial samples for each LMS/LTS-regression. From these prediction errors we take P_{95} and P_{80} for each group. For metabolites in the FACA-category we only use OLS, and present their results for later comparison, giving P_{95}=21.5 s, and P_{80}=14.1 s. We begin with the results for G_L.



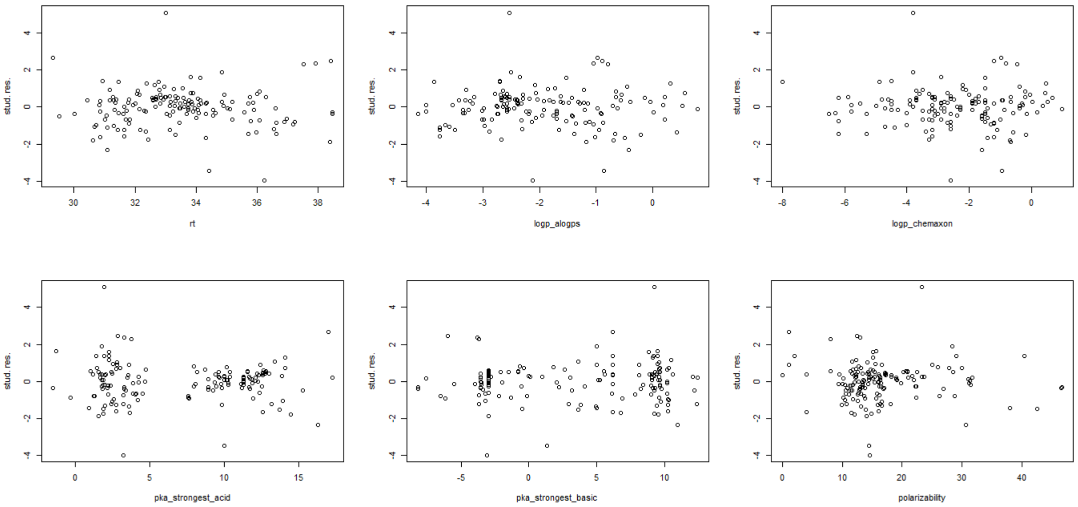
The P_{80}-values are similar across the models, but the P_{95} is smaller for the OLS. The LMS-reweighted OLS seems to make the poorest predictions amongst these models. This suggests we use OLS for G_L-predictions. We continue with G_H.



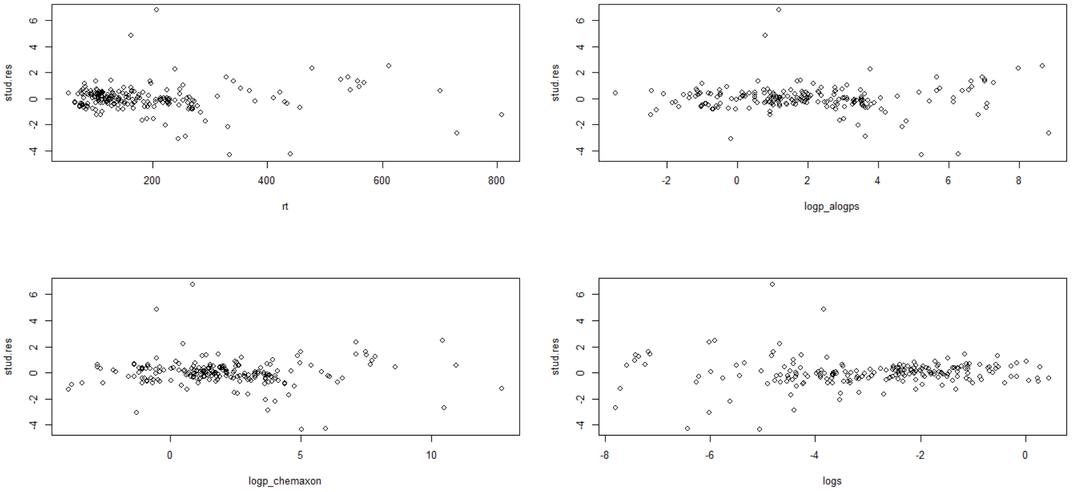
We note that the prediction errors are much larger in G_H than in G_L, which is expected due to the increased variance. The OLS-model outperforms the robust alternatives for P_{80}. In P_{95} the differences are small, which suggests we use OLS for G_H henceforth.

4.4 Model diagnostics

We fit the OLS-model to the whole training data, and present some model diagnostics. We plot the studentized residuals toward the predicted response and the predictors. We begin with $\hat{\boldsymbol{\beta}}_{OLS}(G_L)$.



These plots do not seem to display any serious heteroscedasticity issues. We continue with $\hat{\boldsymbol{\beta}}_{OLS}(G_H)$



From these plots we suggest that there are no heteroscedasticity problems in our models.

Following the quote from Tukey on whether the robust and the classical estimates differ, we also present some coefficient estimates. $\hat{\boldsymbol{\beta}}_{OLS}$ compared to e.g. $\hat{\boldsymbol{\beta}}_{LTS80}$, both fitted on the whole training set, gives the following coefficients.

Table 1: Coefficients of OLS & LTS80 in G_L

Model	Int.	$\log p$_alogps	$\log p$_chemaxon	pka_acid	pka_basic	polariz.
OLS[a]	37.4	0.36	0.71	-0.31	-0.30	0.11
LTS80[b]	36.4	0.38	0.62	-0.19	-0.22	0.09

[a] *Ordinary Least Squares*
[b] *Least Trimmed Squares with 20 % trimming*

In Table 1 we note some small differences in the pka-coefficients. Plotting the pka-predictors against the response suggests that a few outliers are responsible for this increased slope in the nonrobust OLS fit.

Table 2: Coefficients of OLS & LTS80 in G_H

Model	Int.	$\log p$_alogps	$(\log p\text{\_chemaxon})^2$	$(\log s)^2$
OLS[a]	86.5	15.3	4.3	2.2
LTS80[b]	91.1	17.4	2.7	3.1

[a] *Ordinary Least Squares*
[b] *Least Trimmed Squares with 20 % trimming*

In Table 2 we notice that there are some differences between the estimated coefficients, which we will return to in the discussion.

5 Results

In order to get a realistic picture of the prediction error of our model, we now make predictions of our validation set, and present P_{95} and P_{80} for each group in a table, along with the size of each group in the validation set.

Table 3: Sample percentiles of prediction errors

	$FACA^a$	${G_L}^b$	${G_H}^c$
P_{80}	20.3	2.5	46.8
P_{95}	21.5	4.9	86.5
Size	14	80	89

[a] *Fatty Acids and Conjugates, Carboxylic Acids and Derivatives*
[b] *Group with $rt < 45$ s.*
[c] *Group with $rt \geq 45$ s.*

We compare the results in Table 3 with those from the cross validation. The P_{80} of the FACA-class in the validation set is larger than before, but this is probably due to the small number of the FACA-metabolites in the validation set (14 metabolites). In the FACA prediction errors, there are 3 prediction errors in size $\sim$ 20 s, then the error drops to 8 s. Here the training set prediction error $P_{80} = 14$ s based on 45 observations might be more realistic. For the other groups we have more metabolites, and the results suggests small increases in prediction error compared to those from the training set.

6 Discussion

In this study we have compared the predictive accuracy of classical multiple regression and some robust alternatives, to examine whether there was a problem with outliers in the data. This could allow us to predict a majority of the retention times with higher accuracy on the behalf of worse predictions on a small set of outliers. Our results on prediction error indicate that the performance differences between the classical and the robust alternatives are quite small. We've also presented coefficients for some of the models, but comparing these is sometimes complicated by correlation among the predictors. The three predictors in G_H are correlated, i.e. differences in the two $\log p$-coefficients might be balanced by each other, and by the difference in $\log s$, which is negatively correlated with $\log p$. We do believe that the differences in the pka-predictors in G_L are due to some outlying points, but ignoring these does not seem to pay off in prediction error. All in all, the presence of robust methods providing similar results in prediction error is reassuring, since this suggests that there is not a problem with outliers disturbing the OLS-estimates. The robust alternatives have also been helpful in the beginning of the analysis, to identify erroneous database entries.

It is interesting to see that the least trimmed squares regression, here set to trim off 5 % resp. 20 % of the residuals, does not shrink the corresponding percentiles of the absolute prediction errors, compared to OLS. Since the LTS-estimates in higher dimensions are necessarily approximations, the estimates found here might not be accurate enough. Other possible causes is the lack of robustness in the variable selection, small samples and the lower Gaussian efficiency of the LTS compared to OLS. Another possible explanation is that disregarding some percent of the residuals does not necessarily have any effect on the prediction error percentile, unless there is a problem with outliers.

In a previous study of predictions of retention time, Hagiwara et al (2010) noted a linear relationship for higher retention times with $\log p$ and another predictor. This study confirms this finding, and suggests that it can be well explained with only $\log p$ for the metabolite classes "Fatty Acids and Conjugates" and "Carboxylic Acids and Derivatives". Using these classes has cut P_{95} of the prediction error to 1/4 of those from G_H. We believe that incorporating class in retention time predictions could be an important factor in further studies, but this assumes that metabolites in the same class behave similarly. If their characteristics differ much, the class property might be less useful. Constructing a measure of how similar metabolites in a class are might be helpful in deciding when to use class in predictions. However in this study there was not time for this.

Metabolites not well represented is a cause of concern in this study. There might be metabolites which display completely different relationships with the predictors, not represented in this data set. This should be examined from a dataset where the metabolites have been randomly selected.

Another problem which we believe would be helpful to study further is the cause of "jumps" in retention time for some metabolites, when retention time is plotted against $\log p$. In this study we circumvented this problem by splitting the data based on retention time and make two sets of predictions, but there might be more general purposes where a single predicted retention time is of interest.

7 References

Carroll R. J., Ruppert, D., Stefanski L. A., Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, Second Edition. Chapman & Hall.

Chatterjee, S. and Price, B. (1977). *Regression Analysis by Example.*

Wiley series in probability and mathematical statistics

Chen, X., Hong, H., Nekipelov, D. (2007). *Measurement Error Models.* Available from `http://web.stanford.edu/~doubleh/eco273B/survey-jan27chenhandenis-07.pdf` by 2015-05-12.

Croux, C. and Rousseeuw, P. J. (1994). High Breakdown Regression by Minimization of a Scale Estimator. In *Compstat.* Proceedings in Computational Statistics 11th Symposium held in Vienna, Austria, 1994. Editors: Rudolf Dutter, Wilfried Grossman. p. 245-250.

Hagiwara T., Saito S., Ujiie Y., Imai K., Kakuta M., Kadota K., Terada T., Sumikoshi K., Shimizu K., and Nishi, T. (2010). HPLC Retention time prediction for metabolome analysis. In *Bioinformation.* 2010; 5(6): 255258.

Hawkins, D.M. and Olive, D. J. (2002). Inconsistency of Resampling Algorithms for High Breakdown Regression Estimators and a New Algorithm, in *Journal of the American Statistical Association*, 97:457, 136-159, DOI: 10.1198/016214502753479293

Huber, P. J. (1981). *Robust statistics.* Wiley series in probability and mathematical statistics.

Huber, P.J. (2002). *John W Tukeys contributions to robust statistics.* The Annals of Statistics 2002, Vol. 30, No. 6, 16401648.

Liero, H. and Zwanzig, S. (2011). *Introduction to the Theory of Statistical Inference.* Chapman & Hall.

Maronna, R. A., Martin, R.D. and Yohai, V.J. *Robust statistics. Theory and Methods.* Wiley series in probability and statistics.

Rogerson, P. A. (2001), *Statistical methods for geography*, SAGE Publications: London.

Rousseeuw, P. J. (1984). Least Median of Squares Regression in *Journal of the American Statistical Association*, Vol. 79, No. 388: 871-880.

Rousseeuw, P. J. and Leroy A. M. (1987). *Robust Regression and Outlier Detection.* Wiley Series in Probability and Statistics.

Rousseeuw, P. J., and van Zomeren, B. C. (1992), A Comparison of Some Quick Algorithms for Robust Regression in *Computational Statistics and Data Analysis*, 14, 107-116

Rousseeuw, P.J. and Hubert, M. (1996). *Recent Developments in PROGRESS*, Technical Report, University of Antwerp.

Rousseeuw, P. J., van Driessen, K. (2006). Computing LTS Regression for Large Data Sets in *Data Mining and Knowledge Discovery*, 12, 29-45

Sen, A. and Srivastava, M. (1990) *Regression Analysis. Theory, Methods and Applications.* Springer texts in statistics.

Tetko, I. V. Gasteiger, J. Todeschini, R. Mauri, A. Livingstone, D. Ertl, P. Palyulin, V. A. Radchenko, E. V. Zefirov, N. S. Makarenko, A. S.

Tanchuk, V. Y. Prokopenko, V. V. (2005). Virtual computational chemistry laboratory - design and description, in *Journal of Computer Aided Molecule Design* 19, 453-63.