Abstract

Tetrahydrocannabinol (THC), is the primary psychoactive component of cannabis. THC oxidizes into two primary metabolites, 11-hydroxy-9-tetrahydrocannabinol (11-OH-THC) and 11-nor-9-carboxy-Δ-9-tetrahydrocannabinol (THCCOOH), and these three compounds are the primary chemicals detectable in blood plasma. Their concentrations, in relation with time, are responsible for the acute psychoactive effects of marijuana within humans. With the recent legalization and taxation of marijuana use in several states, improved knowledge of cannabinoid metabolism is needed. Studying data from Huestis et al, we notice that the data has distinct nonlinearity. Thus, our intuition suggests that the previously published linear regression models are not the most appropriate method for modeling cannabinoid concentrations. After visual inspection of trends in the dataset, we notice certain "rules" that can be extrapolated to predict time values based on plasma concentrations of THC and its metabolites. This paper proposes a model utilizing random forests, which is a machine learning technique that uses groups of binary trees to model these rules. As a result, this study is able to address some of the shortcomings typically found in linear regression.

Introduction

Motivation

Cannabis is one of the most popular drugs in the United States,[4] and there are increasing concerns how it will affect our driving safety. Experimental data indicates that risk of involvement in a motor vehicle accident increases approximately 2-fold after cannabis ingestion.[5] In order to properly create and enforce drug impairment laws, a better understanding of how cannabinoids metabolize is needed. The primary psychoactive component of marijuana is Δ-9-Tetrahydrocannabinol (THC), which has been shown to form the active metabolite 11-hydroxy–9-tetrahydrocannabinol (11-OH-THC). Further oxidation of 11-OH-THC produces the inactive metabolite, 11-nor-9-carboxy-Δ-9-tetrahydrocannabinol (THCCOOH).[4] Though prior research indicates that 11-OH-THC is considered to be the true active compound, and that THCCOOH is largely inactive with respect to the traditional "high" associated with cannabis.[2]

The main motivation for this paper comes specifically from the passage of Washington Initiative 502, and this paper attempts to relate the average blood concentrations of THC and its metabolites to the level of user impairment. As marijuana legalization continues to garner support across the nation, similar laws to I-502 are likely to increase in popularity, and further understanding of the relation between these psychoactive components and impairment

is necessary.

Introduction to Regression

Let $y \in \mathbb{R}^N$ and $X = [x_1, x_2, \ldots, x_k] \in \mathbb{R}^{N \times k}$. You can **regress** y on X by solving:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

where $\hat{\beta}$ are the coefficients of the regression. The above is equivalent to minimizing the sum of squared distances between y and $\hat{\beta}X \equiv \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$ for all observations:

$$\hat{\beta} \quad = \quad \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \beta X_i)^2$$

where y_i and X_i are the ith rows of y and X, respectively. Once $\hat{\beta}$ is found, the regression model is specified:

$$y = \sum_{i=1}^{k} \beta_i x_i + \epsilon$$

k is the number of variables in the model (columns in the X matrix) while N is the number of observations in the model (rows in the X and y matrices). y is called the **dependent variable**, while $x_1, \ldots, x_k$ are variables or **independent variables.** ϵ is known as the **error** term, which is needed since a straight line does not describe most datasets.

Assumptions for regression. A regression may be bad or incorrectly specified if the following assumptions don't hold:[9]

1. The sample represents the full population.
2. The independent variables do not have measurement error.
3. The independent variables are not linearly dependent.
4. $\epsilon \sim^{iid} \mathcal{N}(0, \sigma_\epsilon^2)$. The errors after fitting the regression are independently and normally distributed, with constant variance equal to $\sigma_\epsilon^2 \in \mathbb{R}$.

Previously Published Model

This paper is concerned solely on the consumption of cannabis through smoking. As a result, the number of currently published models is limited to Huestis' *Blood Cannabinoids II.* Two linear models were derived, which explained the relationships between THCCOOH and/or THC plasma concentrations and the time of exposure.[3]

$$\log(T) = -0.968 \log[THC] + 0.687$$

$$\log(T) = 0.576 \log\left[\frac{THCCOOH}{THC}\right] - 0.176$$

Both models used regression, but with different variable sets. The first model regressed THC plasma concentrations against elapsed time after smoking. The second model regressed plasma THCCOOH/THC ratios versus elapsed time after smoking.[3] In general, the study indicated that the second model was more accurate. We implemented both specifications of Huestis et al. and produced the following output:

```
Call:
lm(formula = log(y) ~ log(THC), data = truncatedSmokerData)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4457 -0.0320  0.3120  0.5117  0.7347

Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept)   1.8550     0.4373   4.242   0.000334 ***
log(THC)     -0.8572     0.1190  -7.206 0.00000032 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9852 on 22 degrees of freedom
Multiple R-squared:  0.7024,    Adjusted R-squared:  0.6889
F-statistic: 51.93 on 1 and 22 DF,  p-value: 0.0000003198
```

Figure 1. Summary of Model I, which regresses THC concentration.

```
Call:
lm(formula = log(y) ~ log(ratio), data = truncatedSmokerData)

Residuals:
     Min       1Q   Median       3Q      Max
-0.96728 -0.16371 -0.00601  0.19408  1.20546

Coefficients:
            Estimate Std. Error t value           Pr(>|t|)
(Intercept) -0.41030    0.08673  -4.731           0.000101 ***
log(ratio)   0.57757    0.02834  20.378 0.0000000000000009 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4051 on 22 degrees of freedom
Multiple R-squared:  0.9497,    Adjusted R-squared:  0.9474
F-statistic: 415.3 on 1 and 22 DF,  p-value: 0.0000000000000009001
```

Figure 2. Summary of Model II, which regresses the ratio of THCCOOH and THC.

Potential weaknesses. Due to the ratio, truncation of zeroes was necessary to avoid dividing by zero. Therefore, using the ratio made applying the full dataset impossible. Additionally, upon implementing the model, we discovered more weaknesses that we will discuss below.

Proposed Model

Data Preparation

The data analyzed in this paper was taken from Huestis' *Blood Cannabinoids I*, where six human subjects smoked a single marijuana cigarette. Their individual plasma levels of THC, 11-OH-THC, and THCCOOH were taken during and after consumption, and they repeated this process for both 1.75% and 3.55% THC cigarettes. In order to simplify the fitting process, We averaged the plasma concentrations for all six individuals. Additionally, research indicates that modern cannabis has become more potent[4], therefore in order to maintain relevancy and better represent cannabis strains being consumed today, this paper focuses only on the 3.55% THC data. Below is a snippet of the dataset used to test our model.

Time (h)	THC (ng/mL)	11-OH-THC (ng/mL)	THC-COOH (ng/mL)
0.017	18.1333	0.0000	0.1667
0.033	48.0000	0.0000	0.2167
0.050	77.8333	0.7500	0.4167
0.067	102.6667	1.7000	1.1700
0.083	115.0000	2.6167	2.0167
0.100	141.6667	3.3667	3.4833
0.117	148.6667	4.6667	5.0500
0.134	144.6667	4.7833	7.1667
0.150	152.0000	5.9333	9.6333
0.167	147.1667	6.3167	12.8333
0.200	126.8333	6.8000	17.3000
0.250	94.8333	7.2333	22.3000
0.300	77.1667	6.3000	29.3333
0.375	48.8333	6.7500	36.3333
0.542	29.6667	4.8500	44.0000
0.792	16.9333	4.2333	44.5000
1.210	9.7167	3.3667	47.5000
1.710	8.0000	2.9167	49.3333
2.210	5.4167	4.7167	48.0000
2.500	4.1167	1.8667	43.5000
4.000	1.8000	1.0500	38.8333
6.000	0.8167	0.2833	26.9000
12.000	0.3833	0.2500	21.0667
24.000	0.0000	0.1667	13.4667
27.000	0.1167	0.1000	13.0667
30.000	0.0000	0.0000	9.7667
36.000	0.0000	0.0000	8.5167
48.000	0.0000	0.0000	7.3167
54.000	0.0000	0.0000	4.9000
60.000	0.0000	0.0000	5.4333
72.000	0.0000	0.0000	4.2333

Figure 3. Average plasma concentrations of THC, 11-OH-THC, and THCCOOH after smoking a single 3.55% THC marijuana cigarette

Introduction to Decision Trees and Random Forests

Overview

We noticed several shortcomings of regression and linear models in general. A regression model can be specified as follows:

$$y_i = \sum_{k=1}^{K} \beta_k x_{i,k} + \epsilon_i$$

where K is the number of variables. This means that for a given variable k, the **marginal effect** of that variable on y is β_k. In other words, a unit increase in variable k leads to an (expected) β_k increase in y_i.

However, our data presents a particular difficulty for regression models. We observe small concentrations of THC and its metabolites for both very small and very large time values. Therefore a single β value will not suffice in capturing this nonlinearity.

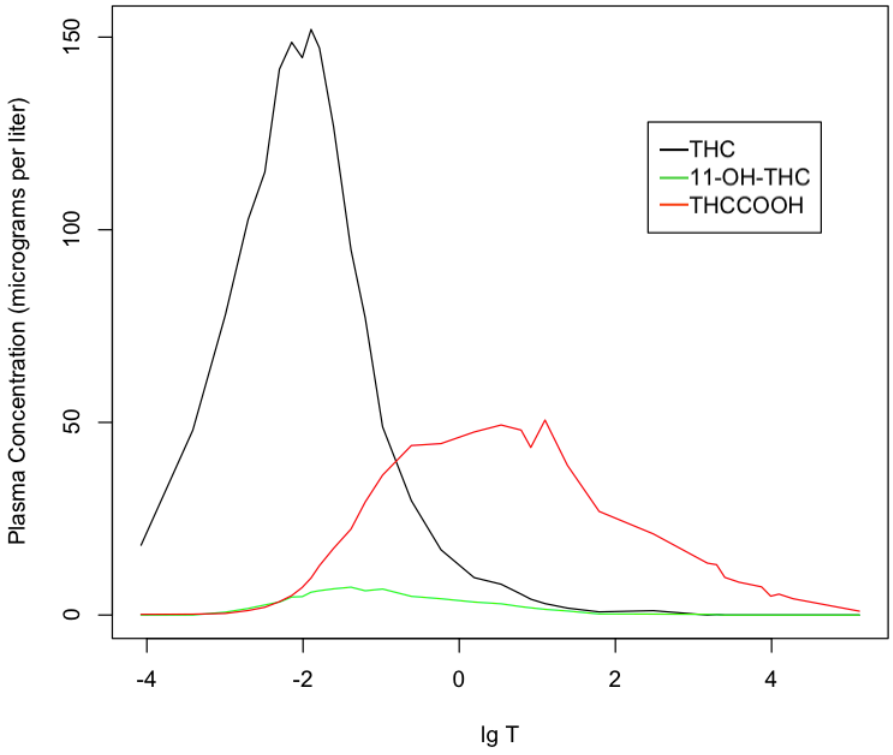

Figure 4. *Plasma concentration (in micrograms per liter) of THC and its cannabinoids. We see that each compound has different peak values and different dissipation rates.*

We chose to use random forests because they can capture "rules", often nonlinear rules, that can solve the above problem. For instance, we can encode intuition as follows:

- When THC levels are high but the other metabolites are low, not much time has elapsed (0 - 0.1 hours has elapsed)
- When THC levels are comparable to THCCOOH and 110H-THC is near its peak, a moderate amount of time has elapsed (0.1 - 0.4 hours has elapsed)
- When THC levels drop below THCCOOH and begin converging to 110H-THC, a large amount of time has elapsed (2.7 hours has elapsed)
- When THC and 110H-THC are near zero, and THCCOOH is much larger but dissipating, a very large amount of time has elapsed (12+ hours has elapsed)

It is difficult for a regression to parse out these "decision rules". The previous study by Huestis et al circumvented these problems by defining the ratio between THCCOOH and THC, which linearly grows with time. However, this requires a certain amount of domain knowledge and also requires truncation of the dataset. As it turns out, we can use random forests to automatically learn these rules.

Decision Trees

A **decision tree** is a binary tree where each node of the tree makes a decision based on a particular variable. The general algorithm for "growing" a tree is as follows:

```
Algorithm 1. Growing a binary tree
while (splits possible):
    for each variable x_k:
        find γ_k for decision rule D_k : x_k < γ_k
    split on rule with maximum entropy
```

The concept of **entropy** is one that is central to decision trees. Intuitively, a decision rule that maximizes entropy "splits" observations cleanly. For example, consider a decision rule with $THC > 0$. This has low entropy because all observations will be partitioned to one side, since concentrations cannot be negative.

A measure of entropy that is commonly used, and the one we will use, is called **Gini impurity:**

$$I_G(f) = 1 - \sum_{i=1}^{m} f_i^2$$

where f_i is the fraction of items labeled with value i in the set.[8] For binary trees, we only have two values of i, namely $I = \{0, 1\}$. When we assess a decision rule $D_k : x_k < \gamma_k$, we use Gini impurity to determine γ_k via a grid search.

Example. Let's say for a particular decision rule, $f_1 = 0.5$ and $f_2 = 0.5$. This means that observations with y values of 0 and 1 are cleanly split by the rule. Then the Gini impurity is $I_G(f_1, f_2) = 1 - (0.5^2 + 0.5^2) = 0.5$. This is the highest that it can be for any decision rule. ■

Example. Let's say $f_1 = 1$ and $f_2 = 0$. Then $I_G(f_1, f_2) = 1 - (1^2 + 0) = 0$. ■

The following is a **regression tree**, fit to our dataset:

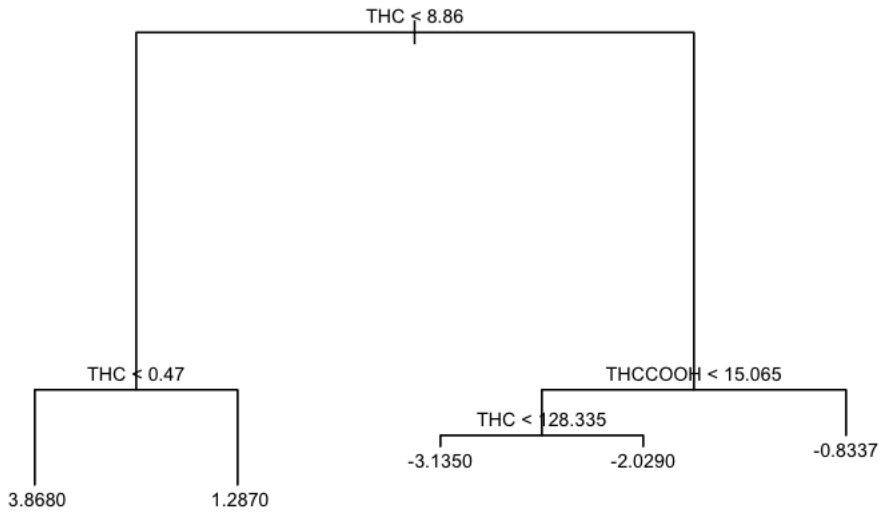

Figure 5. *A binary regression tree.*

On this tree we used $log(T)$ as our dependent variable. We can already see that this tree validates some of our intuitive "decision rules" mentioned earlier. For example, when THC is high and THCCOOH is low, we see that the terminal node output is low, indicating a small T. We provide an example below for further clarity.

Example. Let's say we measured a person to have a THC concentration of $9\mu g/L$ and THCCOOH of $10\mu g/L$. To determine the expected time since smoking for this person, we can use the above tree:

- First traverse right from the root node since $THC = 9\mu g/L > 8.86\mu g/L$, which is the decision criteria at the first (root) node

- Next, we would traverse left, since $THCCOOH = 10\mu g/L < 15.065\mu g/L$
- Finally, we would traverse left, since $THC = 9\mu g/L < 128.335$. This is a terminal node, so it takes the value of $\log(T) = -3.1350$, or $T = 0.04$. So we can say that this individual smoked recently, which is what we would expect since the concentrations are fairly low. ■

Random Forest

We chose not to use decision trees by themselves, but **random forests**, since they allow a combination of decision rules to be incorporated, along with allowing for an overall decrease in model error. [6]

A random forest is a collection of decision trees (sometimes thousands) which are grown in a particular way. The algorithm for random forests is:

```
Algorithm 2. Random forest
for i in 1 to N_trees:
    Take X̃_i ⊂ X, a random sample of the data with replacement
    Grow a decision tree on X̃_i
Average the results of each tree
```

The technique of training a separate decision tree on random subsets of the data is known as **bagging,** or bootstrap aggregation. This allows us to incorporate different decision rules, since if we just grow a tree deterministically, it will always converge to the same tree and thus same rule. Perhaps even more importantly, this allows us to reduce the error of the full model and avoid overfitting. We can see this in the figure below:

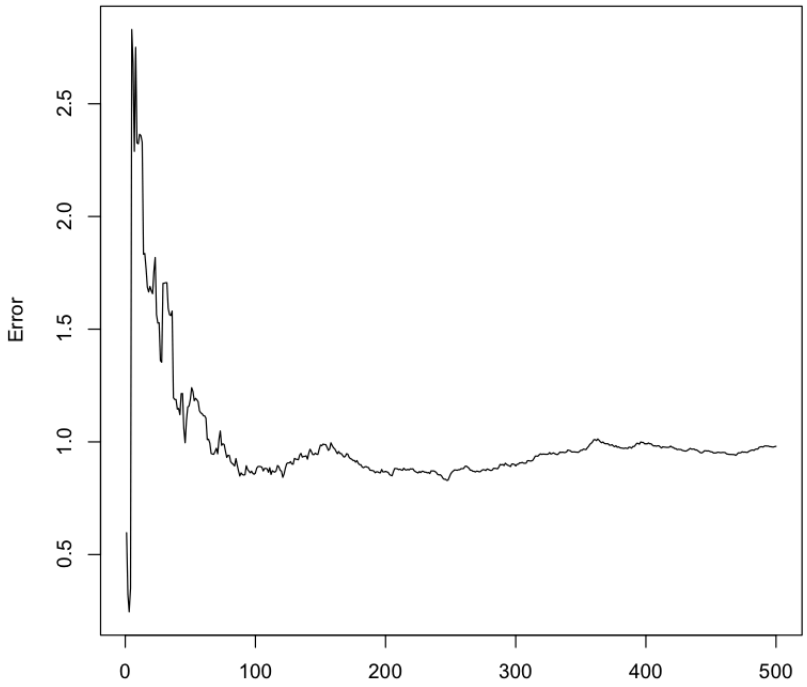

Figure 6. *Number of trees in the random forest vs. mean squared error. We see that the mean squared error converges to about ~1.0 for our model as the number of trees grows to 500.*

We use the mean squared error above, which is defined as

$$\text{MSE} = \frac{\sum_i (\hat{y_i} - y_i)^2}{N}$$

i.e. the average squared error of each prediction. For our model, we can see that it converges asymptotically to some true value $\mu_{\sigma^2} \approx 1$. Once it converges, including more trees does not really help and may actually lead to overfitting. However, including too few trees is also a bad idea, as we can see the large swings in MSE when $N_{tree} < 50$.

In order to obtain predictions from a random forest for a particular observation

$$x_i = (THC_i, THCOOH_i, 11OHTHC_i)$$

we **model average** over all trees in the forest:

$$\mathbb{E}(y_i|x_i) = \frac{\sum_{k=1}^{N_{trees}} g_k(x_i)}{N_{trees}}$$

where $\mathbb{E}(...)$ denotes the expectation operator, and $g_k(x_i)$ is the output for x_i produced by

the kth tree.

Model Details

We use a random forest with variable matrix $X = \{THC, THCCOOH, 11-OH-THC\}$, which is the original set of variables. Because we assume no domain knowledge, this is a natural place to start. We train 100 binary regression trees without **pruning**, i.e. we grow each tree out until the lowest level of the tree consists of nodes that cannot be further split. We chose 100 trees on the basis of the chart above, that the MSE seems to have converged by that time.

We chose to run our model against $\log(T)$ due to the large range of time values within our dataset. The majority of our data is clustered within 0.01-2.5 hours, yet it spans up to 160 hours. By using a logarithmic scale we are able to improve the readability of our data set. Additionally, it seems to improve the model as it unskews the distribution to some extent (not fully, as we shall see).

Results and Comparison

Our random forest model produced the following results:

```
Call:
 randomForest(formula = log(y) ~ THC + THCCOOH + ElevenOHTHC
ee = 100)
               Type of random forest: regression
                     Number of trees: 100
No. of variables tried at each split: 1

          Mean of squared residuals: 1.062396
                    % Var explained: 84.5
```

Figure 7. *Random forest output. We grew 100 regression trees and used one variable per split.*

We see that we "explain" 84.50% of the variance of the model. Turns out this is synonymous with the definition of R-squared in the context of regression. Also known as the coefficient of determination, it is a statistical measure of how close the data set is to the fitted regression line. Below is the equation of R-squared:

$$R^2 = 1 - \left[\frac{\sum(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \right]$$

where N is number of observations. The numerator is the variance explained by the model (sum of squared errors, or residuals) and the denominator is the variance of the dependent variable. This measurement metric conveniently holds for both random forests and regression.

Example. Assume the SSE is

$$\sum(\hat{y}_i - y_i)^2 = 0$$

i.e. we perfectly predict all y_i with our model. Then the $R^2 = 1 - 0 = 1$, which is the highest value it can be. ■

Comparing this R-squared to Huestis et al's model, we see that we beat their first model (their R-squared was 70.24%), but we are short of the second model (94.97%).

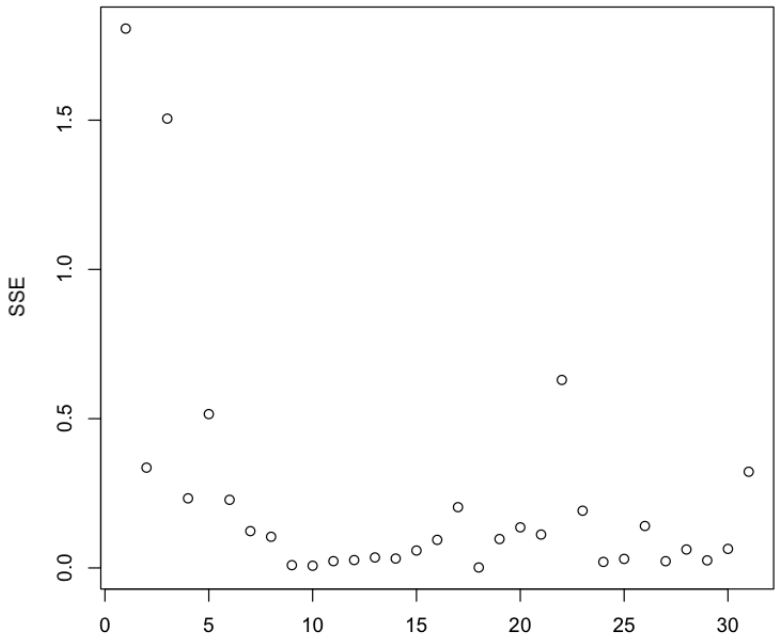

Figure 8. *Residual plot of our random forest model. We see that with the exception of several outliers, the data points are scattered randomly*

For comparison's sake, we also computed the residuals of Huestis et al Model 2 below:

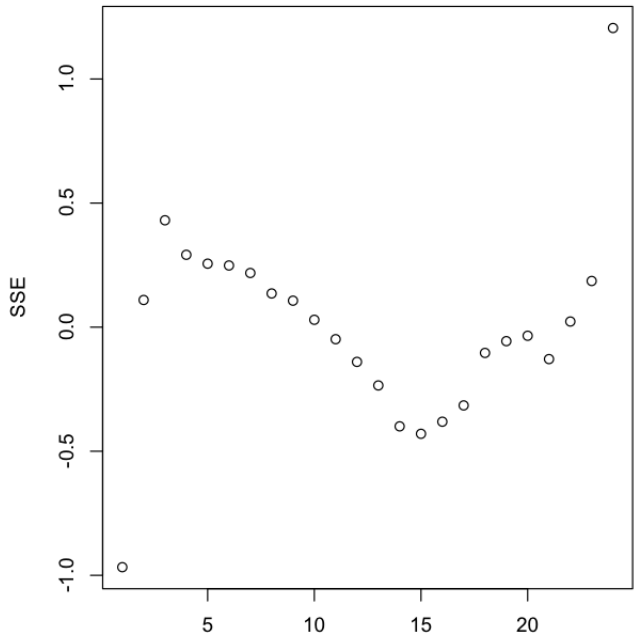

Figure 9. *Residual plot of Model 2 from Huestis et al.*

Note that the residuals in this model violates the regression assumption of independent and perhaps even normal residuals. There is clear correlation between error terms, particularly for middle values of T (seen between indices 4 through 15 above). In comparison, our model's residuals appear much more independent, and apart from a few outliers, could presumably be generated from a normal distribution. Because violating regression assumptions can lead to a biased model, our model may generalize better out of sample to data that it has never seen. From this perspective, our model is an improvement to the literature.

Conclusion

In conclusion, we believe that our model presents several large advantages over linear regression models. By using random forests, we were not forced to make any underlying assumptions, and we were able to use the full dataset without any prior domain knowledge. This generalizes our model, allowing for better representations of new datasets. Our model does have higher error in the tail (i.e. when T $>$ 2.5 hours), but this may be from a lack of substantial data at these times. Overall, we think we made a good effort and in the future, perhaps other more complicated types of trees (i.e. extremely random forests, boosted gradient trees) can be used to improve upon this.

References

[1] Factsheet: Driving under the influence of marijuana, 2013. [Online; accessed 18-November-2014]

[2] Henningfield, J.E., Heustis, M.A., & Cole, E.J. Blood cannabinoids. I. Absorption of THC and formation of 11-OH-THC and THCCOOH during and after smoking marijuana. *Journal of Analytical Toxicology*, 16:276-282, 1992.

[3] Henningfield, J.E., Heustis, M.A., & Cole, E.J. Blood cannabinoids. II. Models for the prediction of time of marijuana exposure from plasma concentrations of delta 9-tetrahydrocannabinol (THC) and 11-nor-9-carboxy-delta-9-tetrahydrocannabinol (THCCOOH). *Journal of Analytical Toxicology*, 16:283-290, 1992.

[4] Huestis, M.A. Human Cannabinoid Pharmacokinetics. *Chem Biodivers.* 2007 August; 4(8): 1770-1804.

[5] Hartman, R.L,. & Huestis, M.A. Cannabis effects on driving skills. *Clin Chem,* 2013 March; 59(3): 478-92.

[6] Breimann, L. Random Forests. *Machine Learning.* 2001 October; 45(1): 5-32.

[7] Breimann, L., & Cutler, A. *Random Forests.* [Online; accessed 18-November-2014]

[8] Decision tree learning. *Wikipedia.* [Online; accessed 22-November-2014]

[9] Nau, R. Regression diagnostics: testing the assumptions of linear regression. *Duke University.* [Online; accessed 22-November-2014]