1^{st} Statistical Report on Breathometer

Linear Model

Thomas Ng

I. INTRODUCTION

This report is written to address the calibration issues for Breathometers. During the production process, plenty of breathometers are produced and labelled by a product ID number(PID). To test whether the breathometers are functioning properly, we developed a statistical selection criteria based on regression, which in turn led us to develop statistical validation procedure to validate the calibration proceudure.The procedures of selection criteria and validation as well as the findings are provided in details in the following sections.

Originally, the selection method proposed involves fitting a straight line between measurement at CP_1 and measurement at CP_4. This is in turn validated by checking whether the measurements at CP_2 and CP_3 lie closely enough to the fitted straight line.

This report also makes comparisons between regression and the original proposed selection method.

II. CALIBRATION (PARAMETER ESTIMATION)

To test whether a single breathometer is working properly or not involves taking voltage measurement from different blood alcohol concentration (*abbreviated as* BAC) levels, i.e. $BAC = 2\%, 4\%, 6\%, 8\%$. Given the voltage measurement (denoted as M in the equation below), we obtain:

$$RLRS = \frac{M}{232 - M} \tag{II.1}$$

And $RLRS$ can be used to deduce its corresponding BAC (to be output or displayed by the breathometer) by

$$BAC = 10^{(\log_{10}(RLRS) - a')/b} \tag{II.2}$$

Equivalently, we can rewrite the above equation such that $\log_{10}(RLRS)$ is linear in $\log_{10}(BAC)$:

$$\log_{10}(RLRS) = a' + b \times \log_{10}(BAC) \tag{II.3}$$

Detailed mathematical derivation of equations (II.1), (II.2),(II.3) can be found in the appendix.

A. Statistical Model - Simple Linear Regression

Given the linear relationship between $\log_{10}(RLRS)$ and $\log_{10}(BAC)$, we propose using a simple linear regression model (II.4) for the measurements obtained at different calibration points, CP_1, CP_2, CP_3, CP_4 (each calibration points correspond to various BAC levels, i.e. $BAC = 2\%, 4\%, 6\%, 8\%$).

$$\log_{10}(RLRS_i) = \alpha + \beta \times \log_{10}(BAC_i) + \varepsilon_i \tag{II.4}$$

where $i = 1, 2, 3, 4, 5, 6$ and ε_i are independent noise with common mean zero and variance σ^2. The α and β can then be estimated by minimising the residual sum of squares (RSS):

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}} \sum_{i=1}^{6} (y_i - \alpha - \beta x_i)^2$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{6} (y_i - \alpha - \beta x_i)^2$$

where $x_i = \log_{10}(BAC_i)$ and $y_i = \log_{10}(RLRS_i)$.

Minimising the residual sum of squares can be viewed as a calibration procedure to reduce noise.

Note that in the calibration procedure, the measurement data for each device is orderd by the timestamp they are collected. The first two calibration points are discarded, as they serve the purpose of warming up the devices and hence, are deemed inaccurate. Specifically, we chose to focus on the devices with 12 data points such that after discarding the first two data points, we can split them into two sets: training set (third data point to eighth data point) and test set (nineth data point to twelfth data point). The dataset given shows that the majority of the devices have 6 data points, which is too little to work with. The second most frequent data points among all devices is 12, which is more reasonable to work with.

Having fitted the regression line, we select those with R^2 greater than 0.9, a positive slope and the p-value of the slope coefficient less than 0.05.

1. Prediction

For prediction, it suffices to compute

$$\hat{x_i} = (y_i - \hat{\alpha})/\hat{\beta}$$

, where $\hat{x_i} = \log_{10}(\hat{BAC_i})$ is the prediction for $log(BAC_i)$.

This is also known as *out-sample prediction*, since we are using the y_i's from the test set, i.e. $9^{th} - 12^{th}$ data points.

B. Standard line Approach

The approach originally proposed involves fitting a straight line between the measurement (y_j) at CP_1 and the measurement (y_i) at CP_4, after discarding the first two data points out of the 12 data points. The slope is estimated by:

$$\tilde{\beta}_{st} = \frac{y_j - y_i}{x_j - x_i}$$

The intercept is estimated by:

$$\tilde{\alpha}_{st} = y_j - \tilde{\beta}_{st} \times x_j$$

Note that negative estimates of the slope coefficient are filtered out here, since the line is expected to be positive.

1. Prediction

For prediction from the standard line approach:

$$\tilde{x_i} = (y_i - \tilde{\alpha}_{st})/\tilde{\beta}_{st}$$

, where $\tilde{x_i} = \log_{10}(\widetilde{BAC_i})$ is the prediction for $log(BAC_i)$. Again, the y_i's are from the $9^{th} - 12^{th}$ data points.

2. Validation

The standard line approach validates its estimated parameters by checking whether the predicted blood alcohol concentration at a particular level $\widetilde{BAC}_i$, which is equivalent to $10^{\tilde{x_i}}$ is within δ of the true blood alcohol concentration level, BAC_i. Specifically, we look at the i's that correspond to samples whose true BAC level at 4% and 6%. δ should be a small value so that the estimated parameters are validated.

$$\text{At } BAC = 4\%, \quad |\widetilde{BAC}_i - BAC| < \delta$$

$$\text{At } BAC = 6\%, \quad |\widetilde{BAC}_i - BAC| < \delta$$

III. FINDINGS

A. Out-Sample Predictions

We use boxplots to display and compare the out-sample predictions between regression and standard line approach. The out-sample predictions for $CP1, CP2, CP3, CP4$, i.e.$BAC = 2\%, 4\%, 6\%, 8\%$ are shown on the next page in figure III.1. Notice that the five boxplot lines correspond to

1. $75th$ quantile $+IQR \times 1.5$, where IQR is the interquartile range and it is equal to $75th$ quantile minus $25th$ quantile.
2. $75th$ quantile
3. median
4. $25th$ quantile
5. $25th$ quantile $-IQR \times 1.5$

Points greater than $75th$ quantile $+ IQR \times 1.5$ or less than $25th$ quantile $- IQR \times 1.5$ are classified as outliers.

There are several points to take away from the boxplots figure.

- There are outliers in both the standard line approach regression at $CP1, CP2, CP3, CP4$.
- The interquartile range in regression is narrower than that in the standard line approach, indicating that the out-sample predictions in regression tend to lie within a closer range of the true blood alcohol concentration (BAC) levels than those in the standard line approach.
- Using regression, the medians of the out-sample predictions are also relatively closer to the true BAC levels than the medians of out-sample predictions from standard line approach.
- Nonetheless, the medians of out-sample predictions from both approaches lie above the true BAC levels. Put simply, more than 50% of the out-sample predictions overestimates the true BAC levels.

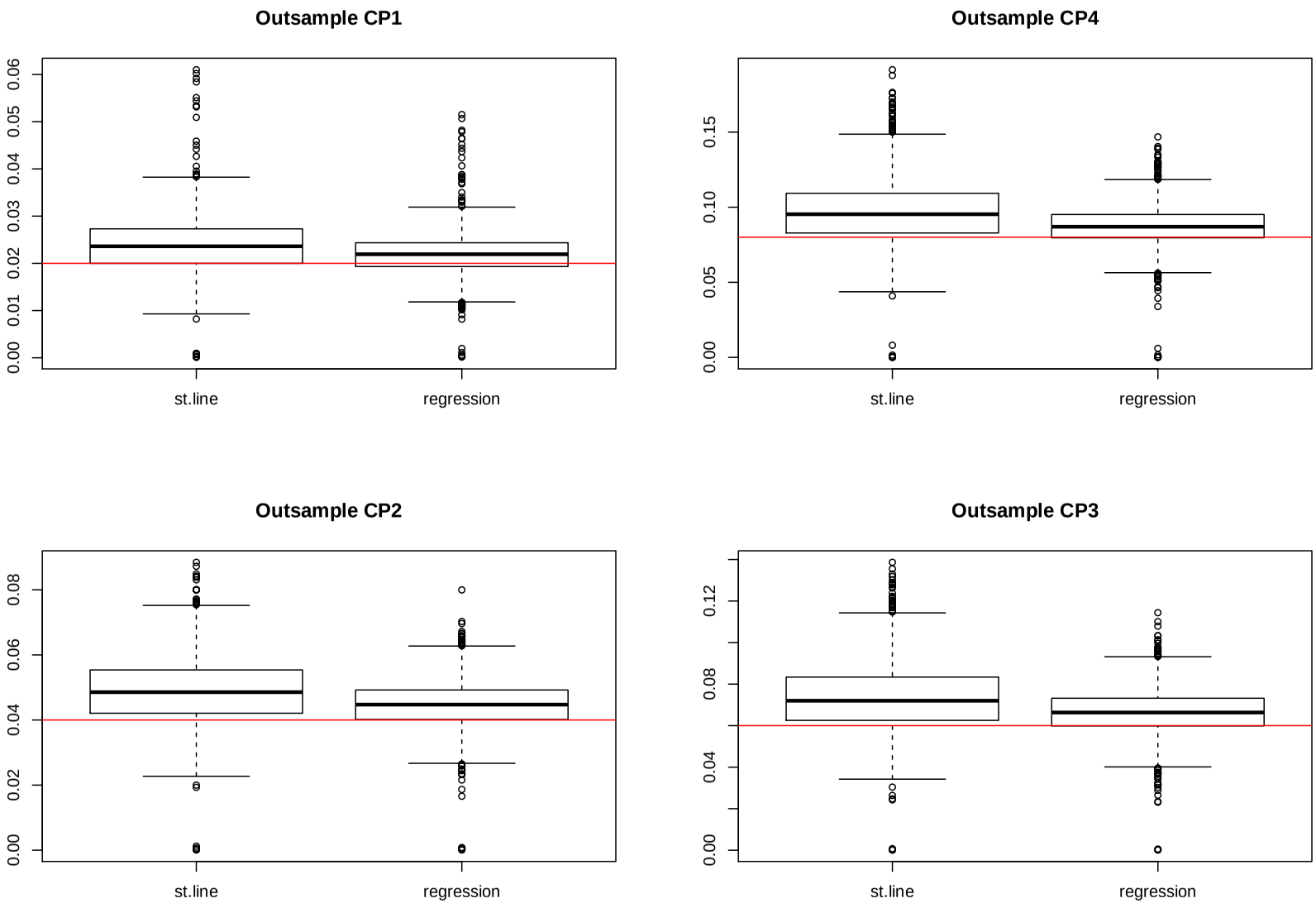

Figure III.1: Out-Sample Predictions

- The average value (*expected value*) of out-sample prediction of $CP1$ is 0.0219456, which indicates an upward bias, and the upward bias is more severe in the case of standard line approach with average value equals to 0.02393132. Similar issue also arises in $CP2, CP3, CP4$.

The *upward bias* issue is something that we will address in the next report.

B. Out-Sample Prediction Error

For each sample and each CP level , the out-sample prediction error in the regression case is computed as follow:

$$d_1 = |\hat{BAC}_i - BAC|$$

And similarly for standard line approach,

$$d_2 = |\widetilde{BAC_i} - BAC|$$

,where BAC is the true blood alcohol concentration level and i corresponds to samples from the test set.

To compare the two approaches, we compute $d_2 - d_1$ to find out which approach has larger prediction error. The boxplots of the difference of the absolute out-sample prediction error between standard line and regression are shown on the next page in figure III.2. It is evident that at all CP levels, the differences are positive (> 0). This illustrates that the standard line approach tends to yield greater out-sample prediction error than regression.

An interesting observation is that the differences increase with the CP levels. This means that as we go from $CP1$ to $CP4$, the difference of the absolute prediction error between the two approaches increases.

C. In-Sample Predictions

Although we care mostly about the out-sample prediction performance of the two approaches, it is worth looking at the in-sample prediction performance because it may explain the puzzles in out-sample predictions. In particular, we look at the $CP1$ and $CP4$

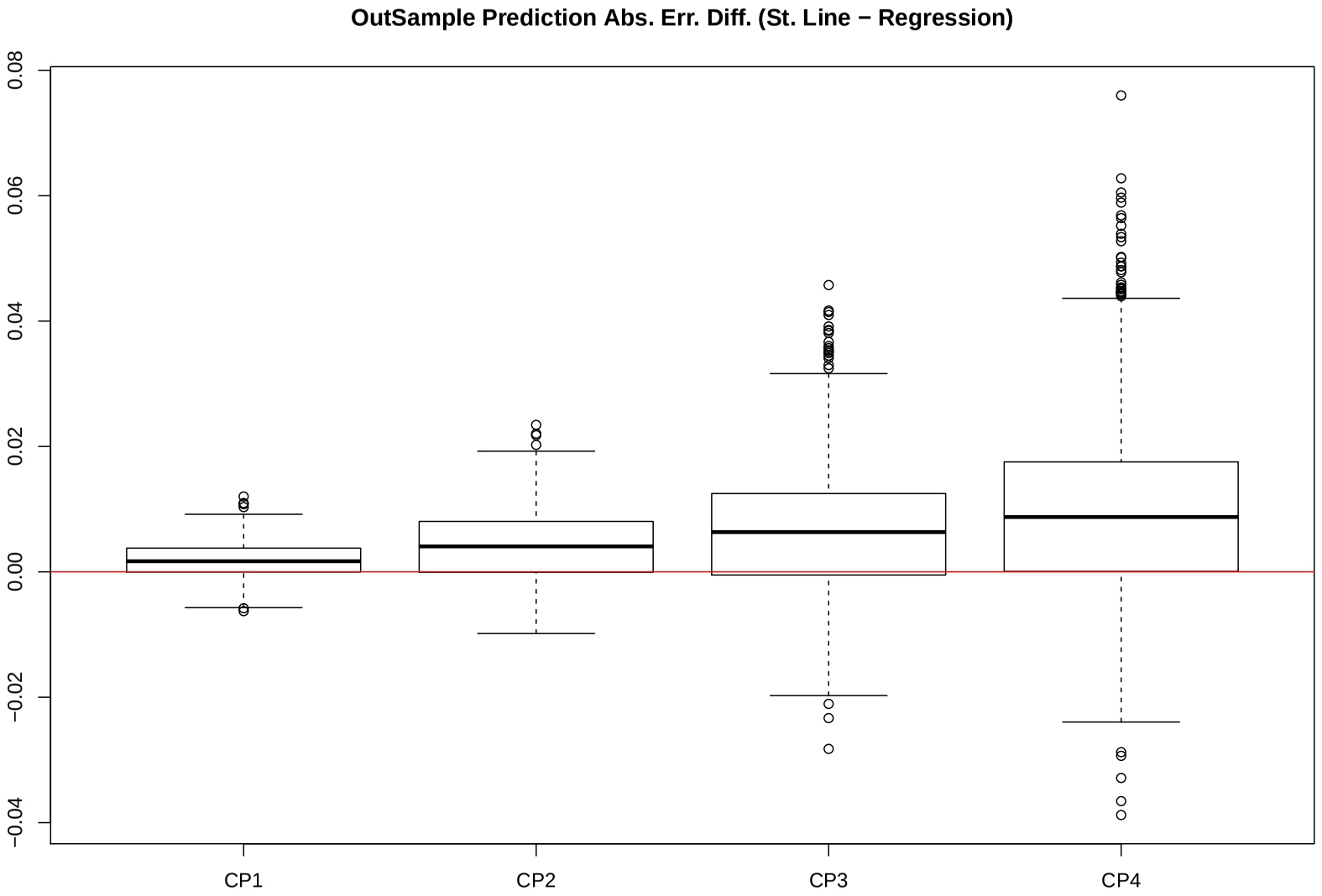

Figure III.2: Absolute Out-Sample Predictions Error - Difference between Standard line and Regression

measurements from the training set.

The corresponding boxplots are shown in figure IV.1. Interestingly, we observe that most in-sample predictions fall below the true BAC level at $CP1$ and $CP4$ in the case of regression.
The in-sample predictions yielded by the standard line approach all lie on the true BAC levels. This is because the standard line approach's parameter estimation procedure only makes use of the two measurements of $CP1$ and $CP4$. So, the in-sample training error will definitely be zero. In the case of regression, there are a few outliers, which is what we would expect since the parameter estimation procedure is also based upon the training set.

IV. CONCLUSION

This leads us to the *bias-variance tradeoff* argument. As standard line approach minimises the in-sample training error, the parameter estimators exhibit very low bias, they face the problem of high variance and this explains why the out-sample predictions from the standard line approach performs worse than those from regression. From another perspective, the regression approach strikes a better (not necessarily optimal) balance between bias and variance than the standard line approach and thus, yields better out-sample predictions.

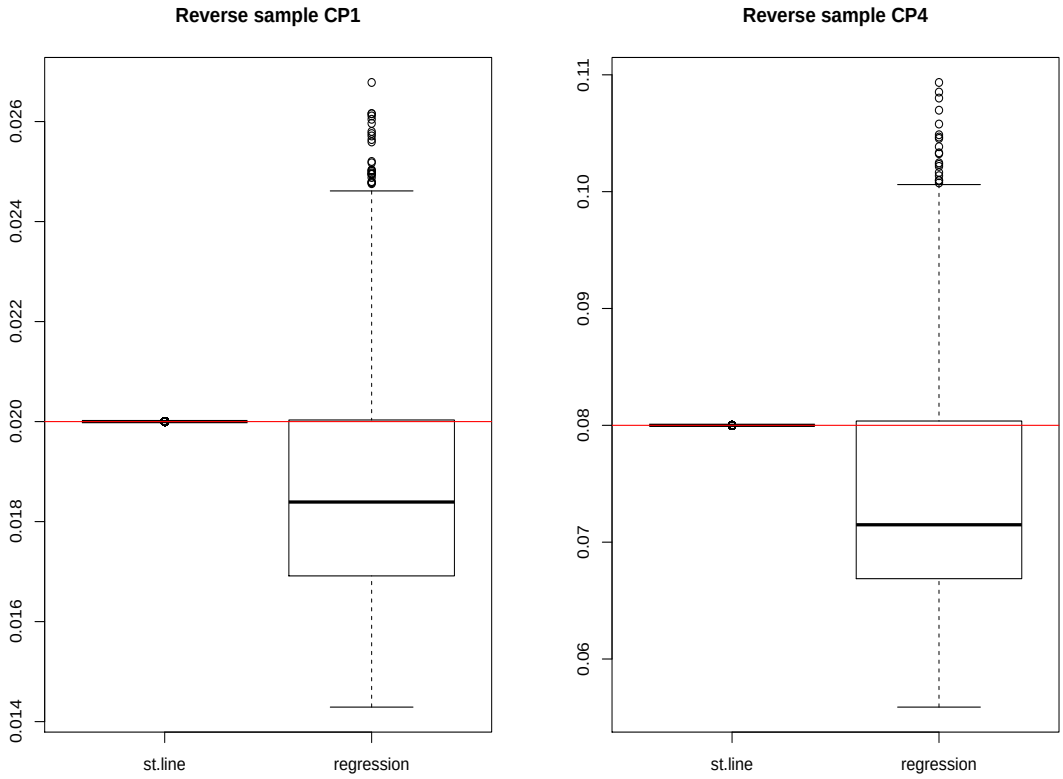

Figure IV.1: In-Sample Predictions - Standard line and Regression

V. APPENDIX

A. Mathematical Model

From the specification of the integrated circuit,

$$V_{RL} = I \times R_L \tag{V.1}$$

$$V_C - \frac{V_H}{2} - V_{RL} = I \times R_S \tag{V.2}$$

where $V_C = 5$ and $V_H = 0.9$ and $0 \leq V_{RL} \leq V_C - \frac{V_H}{2}$. Take (1) $\div$ (2),

$$\frac{V_{RL}}{V_C - \frac{V_H}{2} - V_{RL}} = \frac{R_L}{R_S}$$

R_L is constant ("load" resistance) and $\log_{10} R_S$ is linearly decreasing with $\log_{10}(BAC)$. Let

$$\log_{10}(R_s) = a + b \log_{10} BAC.$$

From the measured V_{RL}, we can then deduce the relationship by using elimination.

$$\log_{10} \frac{V_C - \frac{V_H}{2} - V_{RL}}{V_{RL}} = \log_{10} \frac{R_S}{R_L} = a + b \log_{10} BAC - \log R_L \tag{**}$$

Given we measured at two distinct levels of BAC and obtain two distinct levels of V_{RL}, we have

$$\log_{10} \frac{V_C - \frac{V_H}{2} - V_{RL_1}}{V_{RL_1}} = a + b \log_{10} BAC_1 - \log_{10} R_L \tag{V.3}$$

$$\log_{10} \frac{V_C - \frac{V_H}{2} - V_{RL_2}}{V_{RL_2}} = a + b \log_{10} BAC_2 - \log_{10} R_L \tag{V.4}$$

From equations (3) and (4), one can solve

$$b = \frac{\log_{10} \frac{V_C - \frac{V_H}{2} - V_{RL_2}}{V_{RL_2}} - \log_{10} \frac{V_C - \frac{V_H}{2} - V_{RL_1}}{V_{RL_1}}}{\log_{10}(BAC_2) - \log_{10}(BAC_1)} \tag{V.5}$$

Then, plugging equation (5) back to (3), it is then easy to deduce $a' + b \log_{10} R_L$ and the

formula (**) can be used for future prediction (with $a' = a + \log_{10} R_L$ and

$$a' = \log_{10} \frac{V_C - \frac{V_H}{2} - V_{RL_1}}{V_{RL_1}} - b \log_{10} BAC_2.$$

VI. 8-BIT SCALE

As all voltages are measured in 8-bit scale, it is equivalent to following transformation:

$$\frac{R_L}{R_S} = \frac{V_{RL}}{V_C - \frac{V_H}{2} - V_{RL}} \times \frac{255/5}{255/5} = \frac{M}{232 - M}$$

where M is the measurement in the file and $M = V_{RL} \times 255/5$. $[232 = (5 - 0.45) \times 255/5]$. Using the notation in the file,

$$RLRS = \frac{R_L}{R_S} = \frac{M}{232 - M}$$

The above computation can be rephrased as follows:

$$RLRS_1 = \frac{M_1}{232 - M_1}, BAC_1 = 2\%$$

$$RLRS_2 = \frac{M_2}{232 - M_2}, BAC_2 = 8\%$$

where M_1 and M_2 are the measurements obtained. Then,

$$b = \frac{\log_{10}(RLRS_2) - \log_{10}(RLRS_1)}{\log_{10}(BAC_2) - \log_{10}(BAC_1)}$$

and

$$a' = \log_{10} RLRS_1 - b \log_{10} BAC_2.$$

Then, any other $RLRS$ can be used to deduce its corresponding BAC by

$$\log_{10}(RLRS) = a' + b \times \log_{10}(BAC)$$

or

$$BAC = 10^{(\log_{10}(RLRS) - a')/b}$$