CORE TEMPERATURE REGRESSION MODELLING

Shermaine Si Ying Ying[1], Chan Guangyong Leonard[2], Seah Chun Wei[2]
[1]Raffles Institution, One Raffles Institution Lane, Singapore 575954
[2]DSO National Laboratories, 20 Science Park Drive, Singapore 118230

BACKGROUND AND PURPOSE OF RESEARCH AREA

Around 20 SAF soldiers fall prey to heat injuries every year (Ong, 2014), despite strict prevention measures such as water parades and constant monitoring of weather conditions. Heatstroke, being one of the fatal types of heat injury, occurs when core body temperature is higher than 40^{o}C with a possibility of organ damage and death if the body is not rapidly cooled (Binkley, 2002). It would be beneficial to have a model to predict human core temperature for preventing heat injuries in soldiers engaging in strenuous physical activity; however, human core temperature cannot be obtained easily through non-invasive means such as measuring peripheral skin temperature. Previous research include a simple linear regression model of core temperature from skin temperature (Richmond, 2013) and a Kalman filter time-series (Buller, 2010). This research aims to address this problem by developing a regression model of core body temperature from skin temperature and heart rate, both of which can be measured quantitatively.

HYPOTHESIS OF THE RESEARCH

This project aims to develop a regression model in predicting core temperature given other contemporaneous variables (skin temperature, heart rate).

RESEARCH METHOD AND MATERIALS

Experimental data (denoted as subset B) was obtained from 2 labs which conducted 4 separate experiments on 34 unique subjects, comprising their core temperature (T_{Core}), skin temperature (T_{Skin}) and heart rate (HR) at 5 minute intervals. Non-experimental data (denoted as subset A) comprising age, weight, height, Body Mass Index (BMI) and Body Surface Area (BSA), of which the latter 2 are metrics of body health, was also provided. Both experimental data and non-experimental data were to form the training set, holdout (validation set) and test set with 4311, 944 and 932 points respectively. Both subsets underwent a Box-Cox transformation:

$$x^* = \begin{cases} \dfrac{x^\lambda - 1}{\lambda} & if\ \lambda \neq 0 \\ \log(x) & if\ \lambda = 0 \end{cases}$$

Where x is the original predictor, x^* is the transformed predictor and λ is the optimised parameter indicating the power that all predictor points are raised to, such that the dataset is closer to normality. Skewness in the data is hence minimised, as seen in Figure 1 for the case of HR, where the small increase in density (indicated) is resolved after transformation.

Next, the data was centred and scaled to adjust the variables to zero mean with a common standard deviation of one. Principal Component Analysis was then performed on each subset to find linear combinations of the predictors (PCs) that carried the greatest possible variance.

These PCs were ranked in descending order based on the percentage of variation captured, and were surrogate predictors in determining a smaller set of predictors that could capture most information from the original data.

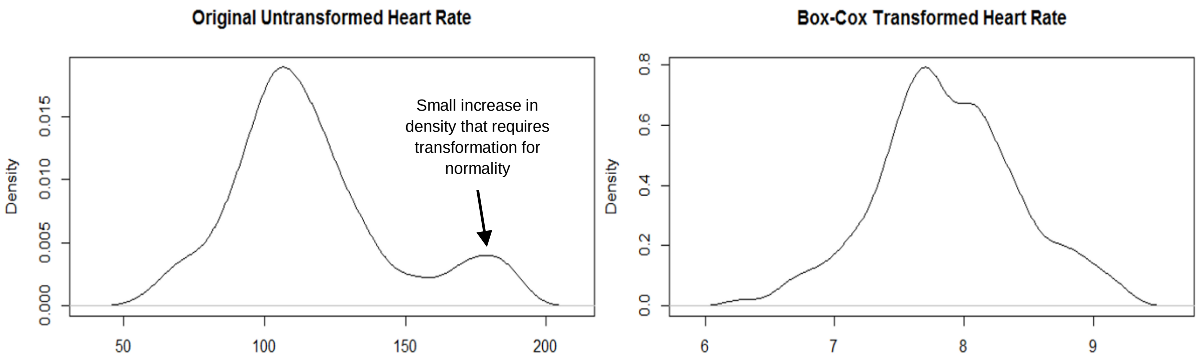

Figure 1: Density plot for Box-Cox transformation of heart rate

After using PCA to determine if any predictors could be removed or combined linearly to reduce the feature size, the following models were considered based on 10-fold cross validation repeated 5 times to reduce bias:

1. Regression
    a. Linear regression
    b. Partial least squares (PLS)
    c. Stepwise
    d. Multivariate adaptive regression splines (MARS)
2. Trees
    a. Classification and regression trees (CART)
    b. Conditional inference trees
    c. Cubist
3. Regularisation
    a. Elastic nets
    b. Least angle regression (LARS)
    c. Least absolute shrinkage and selection operator (LASSO)
4. Neural network
5. Support vector machines (SVM)

The 12 models were evaluated based on 2 performance metrics: root-mean-square-error (RMSE) and the coefficient of determination (R^2). For further comparison based on the holdout data, selected models then underwent iterative model diagnostics for further comparison through residual analysis. The formula for residuals is as follows:

$$\hat{\varepsilon} = y - \hat{y}$$

Where $\hat{\varepsilon}$ is the residual of a point, y is the actual value of T_{Core} in the holdout set and $\hat{y}$ is the predicted value of T_{Core} based on the model used. Adopting a residual approach to analysis provided a well-rounded evaluation of a model's accuracy.

Next, the models were ensembled with the aim of increasing the performance accuracies. 3 types of ensemble models were considered: greedy selection to include the best performing single models with weights, stacking to include a variety of models with unique high-performing aspects, and averaging to weigh the predictions of each model equally. Based on the test set, their performance was evaluated against the best 4 single models (Figure 4) by the following loss functions: mean absolute error (MAE) for absolute accuracy loss, mean square error (MSE) for penalising large residuals, maximal information coefficient (MIC) for correlation, and maximum asymmetry score (MAS) for deviation from monotonicity.

INTERPRETATION OF DATA, RESULTS AND FINDINGS

Preliminary analysis showed that experiment subjects were aged 19 – 24, weighed 52 – 89 kg, stood 1.61 – 1.81 metres tall, had a BSA range of 1.60 – 2.12 m^2 and a BMI range of 18.1 – 28.2. From Figure 2, most participants were relatively healthy (BMI 18.5 – 25) apart from a few outliers. Although the number of test subjects was not large (35), since the model will not be based on a time-series i.e. number of unique points not limited by test subjects' individual characteristics, the data can be modelled after a normal distribution, with 6187 unique points overall. Hence, while a regression model based on this dataset can be robust due to a smaller outlier effect, its accuracy level may vary when predicting T_{Core} for a much older person.

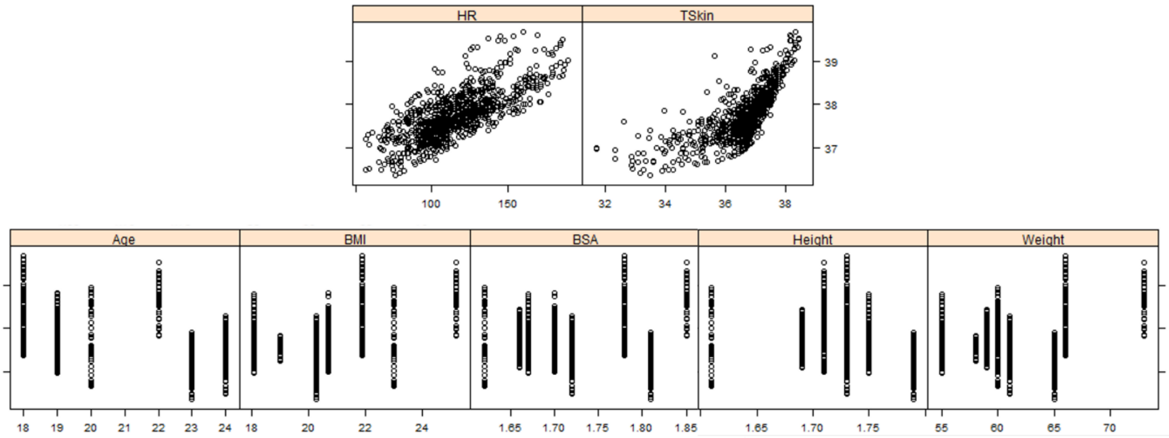

Figure 2: Scatterplots of the predictors in subsets A and B versus T_{Core}

PCA was conducted on the non-experimental predictors in subset A. The largest percentage of the original variability was summarised by PC$_1$ (51.9%) and the smallest by PC$_5$ (0.000290%), which shows that PC$_5$ only accounts for an insignificant amount of variance, PC$_1$ accounts for a slight majority of information in the original data. It would then be reasonable to remove the predictors comprising the majority of PC$_5$ i.e. height and BSA, which occupy greater weight in PC$_5$ as seen in the last column of Table 1, which presents the coefficients of each predictor for individual PCs (rotations) and obtained through the decomposition of the original data matrix. However, these 2 predictors also carried significant component weights in PC$_2$ and PC$_1$ respectively, which comparatively were more representative of the data than PC$_5$. Hence, all 5 predictors in subset A were kept.

Table 1: Individual component weights of predictors in subset A

Predictors	**Rotation Matrix**				
	PC$_1$	PC$_2$	PC$_3$	PC$_4$	PC$_5$
Age	0.3425612	-0.2761554	-0.89767318	-0.02382014	-0.002331995
Weight	0.5829495	0.2496947	0.12536343	0.75130692	0.132832001
Height	0.2894910	**-0.6477958**	0.31266295	-0.16907290	**0.608450320**

Predictors	**Rotation Matrix**				
	PC_1	PC_2	PC_3	PC_4	PC_5
BMI	0.3236982	0.6526855	-0.06423423	-0.53265745	0.425876941
BSA	**0.5951774**	-0.1255110	0.27673581	-0.35022979	**-0.656328622**

PCA analysis conducted on subset B revealed that the 2 measured predictors, T_{Skin} and HR, were correlated. Figure 3 compares the original relationship between T_{Skin} and HR, to that of Principal Component 1 and Principal Component 2, which has undergone a rotational transformation about the axis of greatest variance. Although this seems to show that T_{Skin} and HR used in conjunction measure redundant information and that either predictor or a linear combination of both could replace the original 2 predictors in the model, changes in HR due to the body's physical activity is known to result in a response in T_{Skin} through the human biological system (Cuddy, 2013) so a correlation was expected; since Principal Component 1 accounted for less than 80% of the original variance, neither predictor was removed.

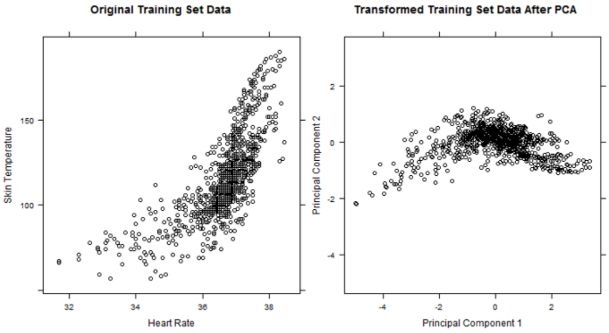

Figure 3: PCA analysis of subset B

The 12 models evaluated based on the training set were evaluated based on 2 performance metrics: root-mean-square-error (RMSE) and the coefficient of determination (R^2). Figure 4 shows the resampling results from across the models, and each of the 50 coloured lines correspond to a common cross-validation. The R^2 value for the LARS and LASSO models are not depicted because they were too large (> 10) and it is evident from the RMSE plot that these models perform poorly overall, while similar greedy models such as stepwise regression also underperformed. This could be due to the increase in bias when these models chose the best set of predictors and considered a larger number of possible models. Conversely, the best performing models (lowest RMSE values and highest R^2 values) are the neural network with 0.1 weight decay and 25 hidden units and Cubist. This could be due to the neural network's optimised ability to detect complex non-linear relationships between predictors while the Cubist model referenced a point's nearest neighbours, as well as previous alternating responses for further accuracy.

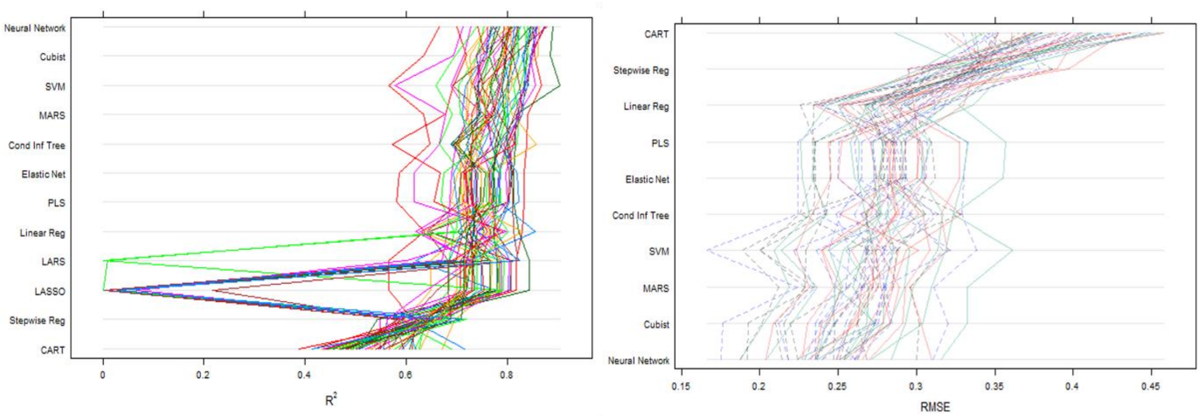

Figure 4: Parallel-coordinate plots denoting performance of models in training set

The 2 best performing single models, neural network and Cubist, both gave similar results when tested on the holdout set. From Figure 5, both models exhibited slight heteroscedascity (non-constant variance) in their residual-fitted plots, as the points are not distributed evenly. However, they satisfy the normal distribution as seen by the linearity of the points on both Q-Q plots, used to compare different quantiles of the same dataset.

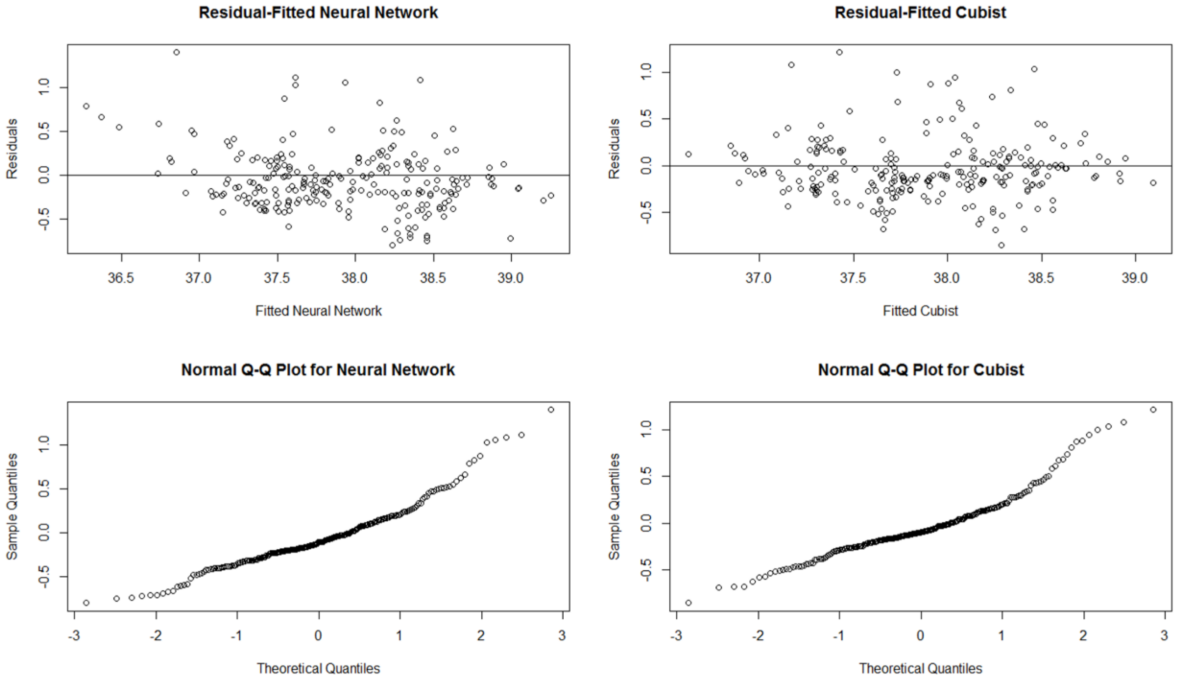

Figure 5: Diagnostics of Neural Network and Cubist models in holdout set

Hence, ensemble models were formed from all single models, excluding LARS and LASSO, and compared against selected best performing singles in the holdout set. Previously underperforming models such as PLS were included in ensemble selection, as their high variance would be reduced in the ensemble models, including these models might be more advantageous.

From Table 2, when loss functions are used as performance metrics, the best performing model is the MARS single model. Although the greedy and stacking ensemble models did not perform too poorly, they still fell short of expectations considering these ensembles selected the best performing single models in the training set to reduce variance and to compensate for individual flaws. The neural network model underperformed, perhaps due to poor choice of initial parameters and overfitting in the training set. The averaged ensemble model performed poorly, as expected, as predictions from weaker models were afforded equal weights as better performing ones.

Table 2: Performance of models in test set (best results for each metric is bolded)

Model		**Performance under loss function performance metrics**			
		MAE	MSE	MIC	MAS
Ensemble	Greedy	0.4469769	0.3033217	0.1890026	0.08656168
	Stacking	0.4034881	0.2400181	0.2302064	0.08007806
	Averaged	2.831272	16.17399	0.1646979	0.09865735
Best single models	NN	1.229479	1.981659	0.1750661	0.06430988
	SVM	0.3881179	0.2295136	**0.2673186**	0.1140114
	Cubist	0.4021212	0.2497234	0.2018318	0.0595927
	MARS	**0.3427453**	**0.1917000**	0.2102119	**0.04886933**

CONCLUSION AND RECOMMENDATIONS FOR FUTURE WORK

The best performing model in the test set was the MARS model. It is important to note that due to real-life issues such as potential health implications from delayed predictions, the final model needs to provide a balance between accuracy and speed. The best performing model may not be the optimal model due to its computational time and complexity. In this case, however, while the stacking ensemble also performed well overall, when exposed to a larger dataset it might prove unfeasible due to the increased quantity of levels to combine data. As a single model, MARS would be preferable for implementation.

Alternatively, predictors could have also been subjected to a spatial transformation which would combine and transform all pre-existing predictors to bring in outliers, resulting in greater normality in the data before carrying out model fitting despite greater difficulty faced afterwards in removing redundant predictor variables, this could be compared to Box-Cox in evaluating model performance afterwards. PCA analysis also revealed that weight was the most important non-experimental predictor in predicting T_{Core}; this could be due to the correlation between one’s body weight and basal metabolic rate, which in turn affects metabolic reactions that produce excess heat during exercise.

In the future, experiments involving heat injuries can be conducted with a classifier for risk of heat stress. This would not only open up the realm of classification models to be trained, but also support the time component as an interval predictor in the model. Models would then reference the duration of physical activity performed and determine thresholds, dependent on how long the activity had taken place, where subjects would be subsequently more at risk of heat injury based on their non-experimental data e.g. BSA. If the physical experiments were able to replicate real-life scenarios, it would be possible for the model to hypothetically determine if one would be at risk of heat injury when undergoing the activity.

REFERENCES

1. Binkley, H., Beckett, J., Casa, D., Kleiner, D., Plummer, P. (2002). National Athletic Trainers' Association Position Statement: Exertional Heat Illnesses. J Athl Train. 2002 Jul-Sep; 37(3): 329–343. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC164365/

2. Buller, M., Tharion, W., Cheuvront, S., Montain, S., Kenefick, R., Castellani., Latzka.,
Roberts, W., Richter, M., Jenkins, O., Hoyt., R. (2013). Estimation of human core temperature from sequential heart rate observations. Physiol. Meas. 34 781. doi:10.1088/0967-3334/34/7/781. Retrieved from http://cs.brown.edu/~mbuller/Buller_InternalTemperatureEstimation.pdf

3. Chow, J. (2011, February 28). SAF is winning the war against heat. The Straits Times.
Retrieved from https://www.healthxchange.com.sg/News/Pages/SAF-winning-war-against-heat.aspx

4. Cuddy, J., Buller, M., Hailes, W., Ruby, B. (2013). Skin temperature and heart rate can be
used to estimate physiological strain during exercise in the heat in a cohort of fit and unfit males. Mil Med. 2013 Jul;178(7):e841-7. doi: 10.7205/MILMED-D-12-00524. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23820362

5. Dietterich, T. (2002, September 4). Ensemble Learning. Oregon State University. Retrieved
from http://www-vis.lbl.gov/~romano/mlgroup/papers/hbtnn-ensemble-learning.pdf

6. Faraway, J. (2002, July). Practical Regression and Anova using R. Retrieved from https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf

7. Filosi, M., Visintainer, R., Albanese, D. (2014). minerva: minerva: Maximal Information-Based
8. Nonparametric Exploration R package for Variable Analysis. R package version 1.4.1.
9. http://CRAN.R-project.org/package=minerva

10. Kenefick, R., Cheuvront, S., Palombo, L., Ely, B., Sawka, M. (2010). Skin temperature modifies the impact of hypohydration on aerobic performance. Journal of Applied Physiology Vol. 109 no. 1, 79-86 doi: 10.1152/japplphysiol.00135.2010. Retrieved from http://jap.physiology.org/content/109/1/79

11. Kuhn, M, Johnson, K. (2014). AppliedPredictiveModeling: Functions and Data Sets for 'Applied Predictive Modeling'. R package version 1.1-6. http://CRAN.R-project.org/package=AppliedPredictiveModeling

12. Kuhn, M. Contributions from Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A.,Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., and Candan, C. (2015). caret: Classification

and Regression Training. R package version 6.0-62. http://CRAN.R-project.org/package=caret

13. Lee, L., Fock, K., Lim, C., Ong, E., Poon, B., Pwee, K., O' Muircheartaigh, C., Seet, B., Tan,
C., Teoh, C. Singapore Armed Forces Medical Corps-Ministry of Health clinical practice guidelines: management of heat injury. Singapore Med J. 2010 Oct; 51(10):831-4; quiz 835. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21103821

14. Mayer, Z., Knowles, J. (2015). caretEnsemble: Ensembles of Caret Models. R package version 1.0.0. http://CRAN.R-project.org/package=caretEnsemble

15. Niedermann, R., Wyss, E., Annaheim, S., Psikuta, A., Davey, S., Rossi, R. (2014).
Prediction of human core body temperature using non-invasive measurement
methods. International Journal of Biometeorology (Impact Factor: 2.1). 06/2013;
58(1). doi: 10.1007/s00484-013-0687-2. Retrieved from
http://www.researchgate.net/publication/237199505_Prediction_of_human_core_body_temperature_using_non-invasive_measurement_methods

16. Ong, H. (June 12, 2014). Keeping it cool. MINDEF. Retrieved from
http://www.mindef.gov.sg/imindef/resourcelibrary/cyberpioneer/topics/articles/features/2014/jun14_fs1.html

17. Richmond, V., Wilkinson, D., Blacker, S., Horner, F., Carter, J., Havenith, G., Rayson, M.
(2013). Insulated skin temperature as a measure of core body temperature for
individuals wearing CBRN protective clothing. Physiol. Meas. 34 1531. doi:10.1088/0967-3334/34/11/1531. Retrieved from https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/13456/3/ES%20insulated%20skin%20v0%206.pdf

18. Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29. http://www.jstatsoft.org/v40/i01/