STA302 Fall 2023 Methods of Data Analysis 1

Final Project Report (Part 3)

Name	Contribution to Proposal
Aaryan	Model Diagnostics; Model selection: Forward/Backward Model Selection; Results, Model Discussion, Ethics Discussion
Anya	Introduction; Assessment of Model Assumptions; Diagnostics; Limitations of Model
Devyani	Model selection: F-test, ANOVA, Correlation and VIF

Introduction

Heart rate variability (HRV) has the possibility of identifying high intervals of stress, or limited physical activity, and suggesting changes to healthier lifestyles (Staff, 2021). That begs the question, when is it indicating an unhealthy lifestyle? If an unhealthy lifestyle is identified, what actions can we take to change it? To answer these, we need to understand what variables influence HRV. For the non-everyday evaluation of health, HRV has shown a correlation with blood-oxygen content, which is useful for high-altitude athletes or travelers in avoiding high-altitude disorders (Saito, 2005). It is also important to specify categorical demographic categories such as gender, as intense exercise impacts “cardiac autonomic functions in women more than men” (Sekiguchi, 2019). Schneider (2018) also studies HRV in athletes with a focus on the intensity of training corresponding to lower HRV. We will investigate these correlations for Fitbit users when multiple predictors are present, with the research question being:

How and to what extent does blood oxygen content, sleep quality, gender (Male or Female), exertion and caloric expenditure affect HRV in Fitbit users?

The final model functional form answers *“how.” “To what extent”* will be explained by their statistical significance and variation explained by the model.

Methods

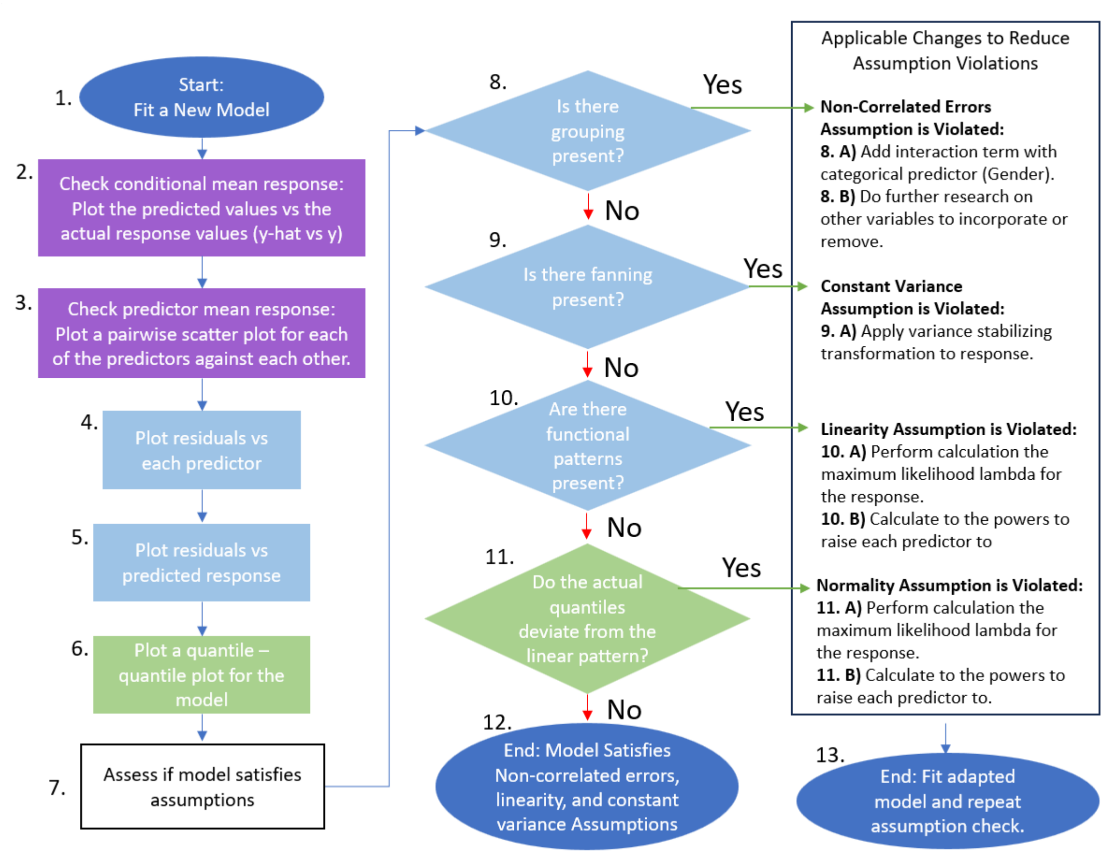
The data was compiled using validated surveys and Fitbit Sense smartwatch daily metrics by the European H2020 Real Time Analytics for the Internet of Sports (RAIS) project [1], collected over 4 months from 71 participants[1]. We obtained this dataset from Kaggle with no project associated.

[1] The European H2020 RAIS project extended from January of 2019 to June 2023, to provide data to train new researchers.

Variable	Expected Correlation to HRV	Description
RMSSD	-	HRV measurement; root mean square of successive heartbeat interval differences; units of milliseconds (ms) (Shaffer, 2017)
SpO2	Positive (Saito, 2005)	Arterial blood oxygen saturation percentage; typical values are between 96% and 99% (Le, 2020)
SPP	Positive (Saito, 2005)	Sleep points percentage (SPP) is the sleep quality index calculated by Fitbit, 1 = best quality of sleep, 0 = worst quality of sleep
EPP	Negative (Schneider, 2018)	Exertion points percentage (EPP) indicates how much one is pushing their body normalized to their maximum level of exertion; 0 = no exertion; 1 = maximum exertion
Calories	Negative (Schneider, 2018)	Calories burned per day; units of kilocalories (unit: calories)
Gender	Negative (Schneider, 2018)	Participant's gender as entered by user; 1 = Male; 0 = Female; Categorical predictor

Figure 1 – Variables selected from RAIS dataset.

Assessing Model Assumptions



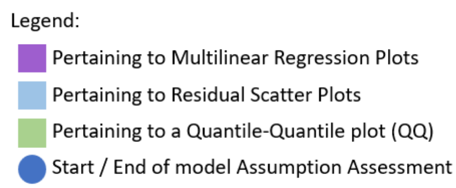


Figure 2 – Process of creating plots, assessing model assumptions, and diagnostic techniques to fix assumption violations.

The decision to create new models was made by assessment of the residual plots mentioned below with aims to improve how the model satisfied the assumptions implicit to linear regression (Figure 2). While fitting our next model, we prioritized fixing the errors identified from the first to the final specified plot (Figure 2).

Specifically, when checking conditional mean response (Figure 2, step 2) we identified any distinguishable functions aside from the linear relationship in the scatter plot. For predictor mean response (Figure 2, step 3), no relationships besides linear relationships occurred between predictors.

The results of the analysis of our assumption and techniques will be discussed with respect to the changes made in the initial model along with the results and an outlier-analysis in our dataset.

Outlier, Leverage Points, and Influential Point Identification

Leverage points were identified through hat matrix diagonal values. Outlier points were identified though standardized residual (Equation 4). Finally, influential observations were identified though Cooks Distance, influence on own fitted values (DFFITS), and influence on beta values (DFBETAS).

Model Selection

In MLR model selection, we assessed the goodness of fit, identified significant predictors, and addressed issues such as multicollinearity. We used the F-test to assess the overall significance of the regression model and compared the fit of the estimated model with other considered models. If the F-test indicated that the model was statistically significant, it suggested that at least one predictor had a non-zero effect on the dependent variable. The partial F-test assessed our understanding of the significance of each predictor while accounting for the presence of other variables in the model, and helped determine if an insignificant variable must remain in the model. We also used Variance Inflation Factors (VIF) to detect multicollinearity, which occurs when predictor variables in the model are highly correlated.

Post checking for multicollinearity, we used likelihood measures (AIC, R-squared etc.) via a series of automated selection tests using the forward (going from 0 to 4 predictors), backward (going from 4 to 0 predictors) and stepwise selection (iteratively adding and removing predictors) to compare Akaike's Information Criterion (AIC) and adjusted R-squared values for models with varying number of predictors. The model that resulted in the lowest AIC for a given number of predictors was chosen as the best for that model for that specific size. Adjusted R-squared values were then used to compare among these to find the most appropriate model.

Results

Table 1 – Variable numerical summaries.

Variable	Mean	Median	Variance	Category Frequency
RMSSD	42	40	0.04	-
SpO2	96	96	1.20	-
SPP	0.79	0.80	0.01	-
EPP	0.76	0.75	0.01	-
Calories	2400	2300	22	-
Gender	-	-	-	477 Male (65%) 211 Female (35%)

Initial Model Assessment

We identified a linear trend that did not follow the expected line through our conditional mean response plot. This indicates possible flaws in interpreting our linear regression assumptions and diagnosing violated assumptions. This will be addressed further with the limitations of our model. No functional patterns were identified in the conditional mean predictor plots. Through our residual plots identified in Figure 2 step 4, 5, these assumptions are not violated; however, there are other concerns including outliers that are also present in the Residuals vs Calories and Residuals vs SpO2.

Finally, the normality of the residuals was also investigated (Figure 2 step 6). Deviation from the norm was observed at the higher and lower quantiles. The adjustments made included (1) performing a log-likelihood estimation on the response, (2) calculation of powers to raise each predictor to (excluding Gender) (Figure 2, step 11. A, 11. B).

Table 2 – Equations obtained from transformation to fix normality.

Model	Eq. No	Equation
Initial Model	1	$HRV \sim SpO_2 + SPP + EPP + Gender + Calories$
Log-Likelihood Adjustment	2	$\sqrt{HRV} \sim SpO_2 + SPP + EPP + Gender + Calories$
Power Transform Adjustment	3	$\sqrt{HRV} \sim SpO_2^{12} + SPP + EPP + Gender + Calories^{\frac{2}{3}}$

Our model summary displayed an F-statistic of 44.68, which is comfortably larger than the 2.98 critical value and lead us to reject the null hypothesis while concluding that at least one of our predictors has a statistically significant relationship with HRV. We further observed statistically significant coefficients for all predictors expect gender and SpO2 and proceeded with partial F-tests for both.

The partial F-test for gender yielded an F-statistic of 47.92, which is comfortably larger than the 3.37 critical value, and thus, we concluded a statistically significant linear relationship between heart rate variability and gender. Therefore, we do not remove gender from our model. We found the same result for our partial F-test for SpO2, which yielded an F-statistic of 58.05.

An analysis of the correlation between our predictors also showed no alarmingly high values. However, since correlation does not consider conditionality of predictors with each other and with the response, we proceeded with VIF calculations. Since all predictors in our model have a VIF ranging from 1.0 to 1.3, we can conclude that this model has some multicollinearity. However, none of the VIF values exceed 5, implying that while mean variances are inflated, this inflation is not alarming in a way that would significantly bias our results. Hence, we continue with our current model regardless of the moderate multicollinearity.

Model selection using likelihood measures (like AIC, R-squared) resulted in all three automated processes (forward, backward, and stepwise) highlighting the model with heart rate variability as the mean response, and Spo2, exertion points percentage, gender, and calories as predictors forming the best model. All three procedures found the AIC to be the lowest (at around 479.5) for this model and the adjusted R-squared to be the highest (at 0.2074). Predictors included in this model were also previously identified as significant when running ANOVA tests on them, further confirming the notion that the chosen model is the best among all alternatives.

Discussion

Model-Based Research Question Conclusions

After performing a series of model assessments, transformations on variables, checking for assumptions, identifying outliers, and iterating through model selection processes the best model we arrived at included heart rate variability (HRV) as the mean response, and SpO2, gender, calories, and exertion points percentage (EPP) as predictors as shown in Equation 6.

Equation 6: $\sqrt{H}RV \sim SpO_2 + EPP + Gender + Calories^{\frac{2}{3}}$

The model shows us that the square root of our response (HRV) is related linearly to the predictors SpO2, EPP, and gender but involves a power transformation of 2/3rds in its relationship to calories. Through tests like ANOVA and F-test, we were also able to identify that all predictors except SpO2 are significant in their relationship to HRV, with calories being the most prominent predictor among them. Further, our tests to identify correlation showed that EPP and gender were positively correlated to HRV. However, SpO2 was negatively correlated, and calories showed a near zero correlation.

Using these results, we formulated an answer to our research question, determining that health and demographic factors like EPP and gender are significant to the HRV of an individual, and that their positive correlation means that an increase in exertion is observed to increase HRV. We also found that this relationship was more noticeable in men as compared to women (and that women's HRV was also less likely to considerably increase upon an increase in exertion).

Although most of these findings agree with our expectations from literature review, we had some particularly surprising results when observing the influence of calories and SpO2 on HRV. We observed SpO2 to have a significant negative effect on HRV, which was contrary to literature where the correlation was expected to be positive (Saito, 2005).

In addition, we observed calories to have a near zero correlation with HRV despite it being the most statistically significant predictor (suggesting a non-linear relationship – Table 2). Contrary to our analysis, literature suggested that this correlation was negative (Schneider, 2018). Lastly, the positive correlation we obtained between EPP and HRV was also not consistent with literature where a negative correlation is seen (Schneider, 2018). We attribute this to the time the measurements were taken, which is continuous in the case of our dataset, but during periods of rest in the research cited.

Limitations of our Model

In our final model, the linear trend did not follow the expected line that was identified in our conditional mean response plot, violating a multilinear regression (MLR) assumption. This indicates an issue both with our resulting model and with the process used to arrive there. The suspected cause was outliers, yet no outliers are identified with this model (Appendix A, Equation 4). However, we do identify multiple leverage points that are visually outside of the main cluster of points in Model Response vs RMSSD plots (Appendix A, Equation 5). This was a concern identified in the first assumption analysis of the model.

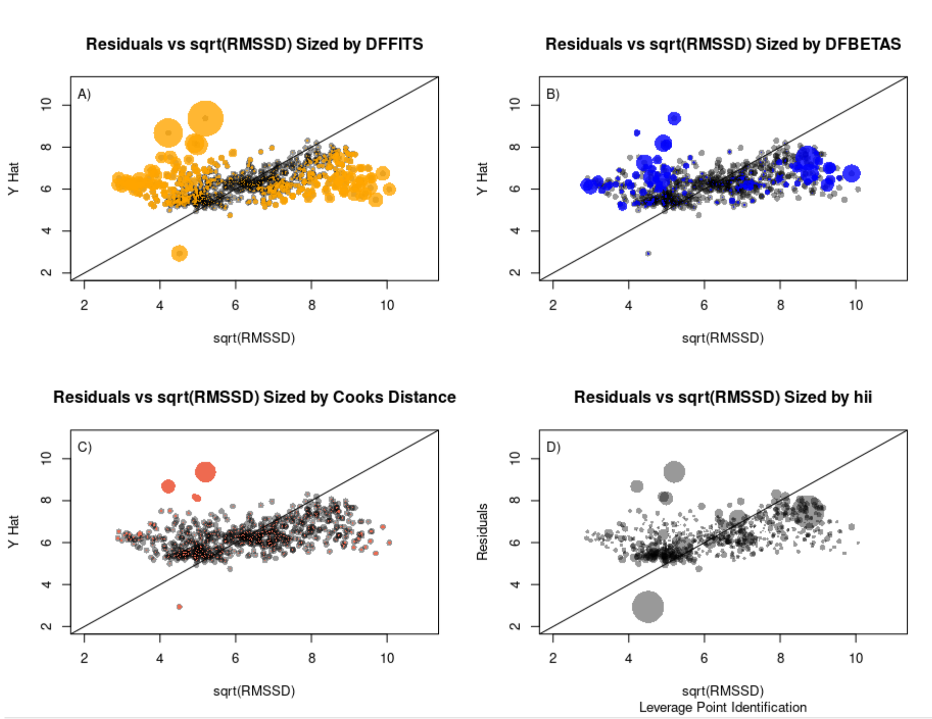


Figure 3 – Influence of points as measured by A) DFFITS, B) DFBETAS, C) Cooks Distance and D) the leverage for each point measured by diagonal values of the hat matrix

The analysis with DFFITS and DFBETAS indicates that these points outside the model have a significant impact on the model results (Figure 3). We do not have a definitive explanation for why select observations fall outside the primary range of values. Possible explanations could include discrepancies in the Fitbits data collection method. Wrist devices tend to have lower accuracy compared to chest-strap monitors when reading HRV (Staff, 2021). Additionally, Fitbit calculates calories consumed based on data that is input by users, possibly reducing the accuracy. Issues with data collection cannot be adjusted by the model.

Another possibility is the absence of critical predictors in the model, resulting in omitted variable bias. This would explain the low R-squared observed for each of our fitted models. If true, it would indicate that our model violates a linear regression assumption, and no widespread conclusion can be drawn. However, we can still comment on how the selected predictors correlate with HRV as observed for this data set.

Conclusion

For our data set, the relationship of each predictor with HRV is presented in the functional form of the final model (Equation 6).

The most significant relationship was observed with calories burned (Appendix B, Table 3). Furthermore, HRV was observed to increase with increases in exertion level, a result that differed across different genders. There was an insignificant relationship between HRV and sleep levels, indicating that sleep levels likely had no influence on HRV for Fitbit users. A lack of statistical significance was also identified for blood oxygen content. Thus, while we cannot make conclusive statements (as our model accounts for around 20% of the variation seen in the data set), this report provides a foundation for data analysis of variables measured by a Fitbit and their relationship with HRV.

Ethics Discussion

The model selection method implemented in this project was the automated method. This was done solely due to the substantial number of predictors in our models, and to save on time that would have been spent manually obtaining likelihood measures for each one of them. It would not be fair to assume that both methods are completely identical ethically since the automated method does not compare every possible model like the manual one does. Secondly, the automated method only utilizes AIC as a primary factor to make decisions regarding including or omitting predictors, even if other likelihood measures (BIC, RSS etc.) disagreed with the results produced by it.

A big ethical factor that the automated method overlooks is the lack of checking assumptions before performing model selection. This means that the best model selected could still violate assumptions and might not be suitable for predictions (giving results that might not be true based on what is observed). To overcome these ethical shortcomings, we took certain mitigating steps. Since the automated methods in our analysis all yielded the same best model, we performed in depth model diagnostics and assumption checks on this model and ensured that the likelihood measures for this model agreed with each other. Through the results obtained from this strategy, we were able to confirm that our model is free of any violations and optimal for use.

Word Count: 1975

References

Le, T. (2020, July 23). *What is oxygen saturation (SpO2)?* iHealth Labs Inc. https://ihealthlabs.com/blogs/faq/what-is-oxygen-saturation-spo2-what-is-the-normal-range-for-spo2#:~:text=Oxygen%20saturation%20(SpO2)%20is%20a,normal%20for%20a%20given%20individual

Saito, S., Tanobe, K., Yamada, M., & Nishihara, F. (2005). Relationship between arterial oxygen saturation and heart rate variability at high altitudes. *The American Journal of Emergency Medicine*, *23*(1), 8–12. https://doi.org/10.1016/j.ajem.2004.09.023

Schneider, C., Hanakam, F., Wiewelhove, T., Döweling, A., Kellmann, M., Meyer, T., Pfeiffer, M., & Ferrauti, A. (2018). Heart rate monitoring in team sports—a conceptual framework for contextualizing heart rate measures for training and Recovery Prescription. *Frontiers in Physiology*, *9*. https://doi.org/10.3389/fphys.2018.00639

Sekiguchi, Y., Adams, W. M., Benjamin, C. L., Curtis, R. M., Giersch, G. E. W., & Casa, D. J. (2019). Relationships between resting heart rate, heart rate variability and sleep characteristics among female collegiate cross-country athletes. *Journal of Sleep Research*, *28*(6). https://doi.org/10.1111/jsr.12836

Shaffer, F., & Ginsberg, J. P. (2017, September 28). *An overview of heart rate variability metrics and norms*. Frontiers in public health. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5624990/

Staff, H. H. P. (2021, December 1). *Heart rate variability: How it might indicate well-being*. Harvard Health. https://www.health.harvard.edu/blog/heart-rate-variability-new-way-track-well-2017112212789

Appendix A: Equations

Equation 4: Normalized residual condition to identify outlier points.

$r_i = \frac{e_{hat\,i}}{s\sqrt{1-h_{ii}}}$;

$$r_i \notin [-4,\ 4]\ if\ n \geq 50$$

h_{ii} - diagonal Hat matrix value for row i

$e_{hat\,i}$- estimated model residual, for observation I

s - square root of sample standard of deviation

r_i - Standardized residual

n - Total number of observations

Equation 5: Diagonal Hat condition to identify leverage points.

$$h_{ii} > 2\left(\frac{p+1}{n}\right)$$

h_{ii} - - diagonal Hat matrix value for row i

p - -Number of Predictors

n - -Total number of observations

Appendix B: Tables and Figures

Table 3: Regression Analysis Results for Selected Model

Variable	Coefficient (Standard Error)	T-statistic	P-value
Intercept	7.363 (4.847)	1.519	0.129
SpO2	-0.057 (0.050)	-1.134	0.257
Gender	0.236 (0.129)	1.816	0.069 *
Calories Burned	0.017 (0.002)	9.827	0.000 ***
Exertion Points Percentage	1.431 (0.484)	2.954	0.003 ***

Note: (*) for significance on 10%, (**) for significance on 5%, (***) for significance on 1%