Does our exercise affect how well we sleep?

Jonas Freeman
Harvard University
Dept. of Statistics

Dries Rooryck
Harvard University
Dept. of Statistics

Eva Harris
Harvard University
Dept. of Statistics

Caden Woodall
Harvard University
Dept. of Statistics

Sam Lindemann
Harvard University
Dept. of Statistics

1 Introduction and Motivation

1.1 Motivation

Being healthy is more than a mindset; it is a lifestyle, and people commit to this lifestyle to varying degrees. From elite athletes to the "gym-bro," everyone who takes part in this lifestyle strives for optimal results. The athletic mindset will take you far, but you will only see an impact by training the body. It is impossible to see from the outside of your body exactly what is happening internally. In the present day, even 1% gains are celebrated achievements among professional and Olympic-level athletes. **But if you are already performing incredibly well, how do you become the best that your physiology will allow? This is where WHOOP comes in.**

WHOOP is one of the best fitness tracking bands in the world. All of the data is collected continuously over time, and the app interface uses extensive visuals to display this data. Its founder, Will Ahmed, is a Harvard alum who used his time at school to research the body's response to cardiovascular load and physiological strain. He believed that there was a better way to approach training geared toward maximizing performance. The band syncs roughly 100 times per second, creating a continuous dataset that provides valuable insights to allow the user to optimize their training. [2]

In a 2016 study, the Activity Strain metric of WHOOP was proven to better predict the performance of MLB pitchers than counting the number of pitches they threw (the standard practice at the time for MLB). WHOOP's Activity Strain is a measure of the intensity and duration of physical exertion, expressed on a scale of 0 to 21. It primarily considers cardiovascular load, heart rate zones, and the duration of the activity to derive a composite score. In a piece written by Will Ahmed himself, titled "WHOOP Approved for In-Game Use in Major League Baseball," WHOOP advertises that following this study, it was the first continuous tracking device of its kind to be allowed for use in an MLB game. As an example of its application, the piece illustrated the following scenario: "Clubs could find out that Player X is no longer effective when his Day Strain hits a certain number, regardless of whether he's thrown 50 pitches or 100 pitches to that point." [1]

However, there is more to predicting performance than Activity Strain. For instance, while optimizing the load on your muscles is a critical way of looking at the game, what about the recovery leading into the game? Every day, muscular structures degrade at a cellular level, and every night, they are rebuilt. This process generally occurs during restorative sleep. Restorative sleep is the combination of deep sleep (SWS) and REM sleep. Deep sleep (SWS) is known for the physical restoration that occurs throughout the body overnight. REM sleep is known as the mentally restorative stage of sleep, during which short-term memories are converted into long-term memories. [5]

1.2 Introduction

In this project, we will address these findings by attempting to use other athletic parameters from real WHOOP data to predict the amount of time spent in deep sleep relative to overall sleep. We hypothesize that there is a relationship between this ratio and the load that workouts (or strain) place on the body. We want to find out what the interesting relationships are that we can decompose, and we want to be able to predict consistently based on a person's exercise activity and some limited data on their sleep, what their sleep quality might look like.

2 Data & EDA

2.1 Data Collection & Cleaning

This data was collected via the sensors on Eva's WHOOP band and synced to the WHOOP interface. We were then able to export the data in CSV format. Originally, data was provided in three categories: **Workouts.csv** (covering details like workout duration, activity type, strain, energy burned, and heart rate zones), **Sleeps.csv** (including sleep cycle durations, efficiency, debt, and consistency), and **Physiological_Cycles.csv** (capturing recovery, resting heart rate, skin temperature, blood oxygen levels, and strain).

We chose to take a time-series approach. In the context of time series analysis, each day represents a single observation, allowing for the systematic tracking of workout metrics over time. We decided to create a dataframe that covered the days that Eva had a workout logged (and would be accounted for in workouts.csv), a dataframe that gathered the days that she did not have a workout logged, and then combine the two so that we have one, comprehensive dataset. There was significant redundancy in the data, which required extensive cleaning, so for simplicity, we decided to only focus on `workouts.csv` and `sleeps.csv` for this project. After cleaning, we had **one wrangled dataset** that only included data from `workouts.csv` and `sleeps.csv`. This final wrangled dataset contained 33 variables.

Some common issues observed were tracking errors, such as incorrect activity detection and fragmented sleep data. Fortunately, most of these errors could be corrected manually through the app before exporting the data into CSV format.

2.2 Preparing the Exploratory Variables

We first filtered days that included workouts based on `workouts.csv`. Columns with no data or redundant information were dropped, and the Activity.name variable was converted to a factor for categorical analysis. For clarity and interpretability, we set a floor value on the lowest number of observations of a workout. Activities labeled as "Other" (with only one observation) and "Activity" (with 65 observations) were removed. The Workout.start.time was converted to a Date type, and only the date portion was extracted to aggregate workout metrics by day.

Each of these daily workout summaries included several key metrics: `Total Duration`, which represents the sum of workout durations for each day, and `Total Energy Burned`, which is the total calories burned during workouts on that day. To account for workout duration when assessing heart rate, a `Weighted Average Heart Rate` was calculated, giving more weight to longer workouts. Additionally, weighted averages for each heart rate zone (1 through 5) were derived to capture the distribution of time spent in different heart rate intensities, again weighted by workout duration. By weighting the time spent in each heart rate zone by workout duration, you account for how long you maintained specific intensity levels across different workouts. This helps balance the influence of longer workouts on the daily average, which intuitively makes sense to impact the need for Deep SWS sleep. The `Max Heart Rate` was recorded as the highest heart rate for each day, reflecting peak exertion levels.

To gain a comprehensive view of workout types, the dataset includes counts for each activity, where each activity type (e.g., running, cycling) is tallied based on boolean indicators (0 or 1). A new metric, `Heart`

`Rate Zone 0`, was introduced to quantify the percentage of time spent in a resting or low-intensity state. This was calculated by subtracting the sum of the percentages of heart rate zones 1 through 5 from 100, with the result rounded to two decimal places. These daily aggregates form a robust time series dataset, facilitating the analysis of workout trends, intensity distributions, and recovery patterns over time. This structured approach to summarizing daily workout data ensures that each observation captures a comprehensive snapshot of physical activity, allowing for meaningful insights into performance and recovery.

The other part to time-series analysis is ensuring a complete and continuous date range. We took additional steps to ensure that this assumption was held. The process begins by generating a sequence of dates from the earliest workout date to the latest workout date in the `daily_workouts` dataset. This ensures that no dates are skipped, providing a consistent timeline.

We then used a new dataframe to represent rest days. For each date in the range, default values are assigned to indicate no physical activity and resting heart rate conditions. Metrics like `Total Duration` and `Total Energy Burned` are set to 0, reflecting no workout activity. Heart rate-related variables, such as `Weighted Average Heart Rate` and `Max Heart Rate`, are assigned a resting value of `57 bpm`. Additionally, the weighted average heart rate zones (Zones 1 through 5) are set to 0%, while `Heart Rate Zone 0` is set to 100%, indicating that all heart rate activity is in the resting range. The activity count variables (e.g., Total Running, Total Cycling) are also set to 0, indicating no activity was performed on these days. By incorporating these rest days into the dataset, the continuity of the time series is maintained, which is crucial for generating accurate insights and visualizations.

By combining the dataframe of rest days with the original data from workouts.csv, we were able to make sure that the days with recorded workouts retain their actual metrics, such as higher maximum heart rates and real weighted heart rate zones. Our use of default placeholders in the dataframe of rest days did not overwrite or exclude higher heart rate values in the original dataset, but instead allowed for a clear distinction between workout days and rest days. After merging, the dataset was processed to remove duplicate entries, which kept only the first occurrence of a workout that was logged. To maintain the integrity of the time series, the dataset was then sorted chronologically.

2.3 Preparing the Response Variable

We removed `Cycle.start.time`, `Cycle.end.time`, `Cycle.timezone` because these are time-related metrics that are not in line with our goal. Our goal is to predict the ratio of time spent in Deep (SWS) Sleep to the overall amount of time spent asleep. We plotted a correlation matrix of sleep.csv to understand how the variables relate to one another better and to justify other changes to the dataset. In this, we found that there were four highly correlated variables.

Figure 1 shows the correlation matrix for the sleep dataset.

The correlation analysis of the sleep dataset was conducted using a threshold of 0.85 to identify highly correlated data. The resulting correlation graph revealed a strong correlation between four variable pairs: asleep duration and sleep performance, in-bed duration and sleep performance, in-bed duration and asleep duration, and sleep efficiency and awake duration. To avoid issues of collinearity, we dropped in-bed duration and sleep performance because asleep duration proved to be a reliable empirical measure. Additionally, while not caught by the threshold, sleep debt and sleep need appear to be highly correlated in the graph above. We decided to drop sleep debt as a result.

In-bed duration refers to the total time spent in bed, whether asleep or awake. Sleep performance is a metric calculated by a proprietary WHOOP algorithm that takes into account how well you were able to meet your body's need for sleep, incorporating sleep debt and naps. Being able to reproduce these values simply with asleep duration simplifies things significantly. Similarly, we also dropped sleep efficiency because awake duration proved to be a reliable empirical measure. Sleep efficiency is a metric calculated by a proprietary WHOOP algorithm to measure how fast you go to sleep.

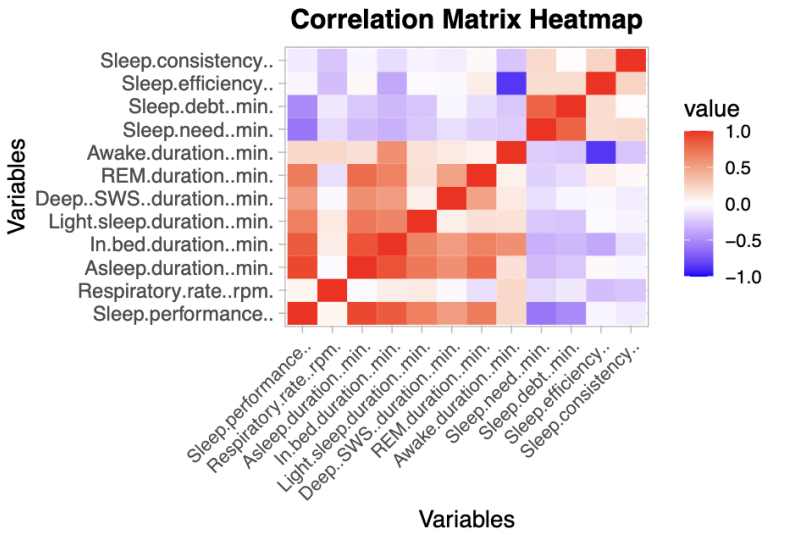

Figure 1: Correlation Matrix of the Sleep Dataset.

Nap statistics were then calculated by converting `Wake.onset` to `Wake.date` (date-only format) and grouping by `Wake.date`. For each date, `Nap.count` was computed as the number of distinct wake-up times minus one, and `Nap.duration` was the total sleep duration minus the longest sleep (main sleep). If no naps occurred, `Nap.duration` was set to zero. This provided clear daily measures of nap frequency and duration.

The main sleep session for each day was identified as the entry with the maximum `Asleep.duration..min.`, retaining the first occurrence in case of ties. To align sleep with workout data, a `Sleep.date` column was created by subtracting one day from `Wake.onset`, reflecting the night the user went to sleep. The main sleep data was merged with nap data using a left join on `Sleep.date` and `Wake.date`, filtering out rows with missing nap statistics. The sleep data was grouped by the wake-up date to organize sleep data by the day the user woke up. The final version of **Sleeps.csv** contained primary metrics for sleep duration, nap count, and total nap duration, and all aligned by date for consistency with the cleaned **Workouts.csv** data.

2.4 Combining Exploratory and Response Data

Next, we used the date of each observation to perform an inner join that combines the dataset that contained our exploratory variables with the dataset that contained our response variables. Additionally, to streamline this dataset, time-related columns such as `Sleep.onset`, `Wake.onset`, and `as.Date(Wake.onset)` were removed, as well as the count for each type of workout. The number of times Eva logged each type of workout is extraneous because it is the heart rate zones that trains the cardiovascular system, irrespective of the actual exercise or method used to achieve this heart rate.

2.5 Our final wrangled dataset

The final dataset included the following columns: Respiratory rate (rpm), asleep duration (min), light sleep duration (min), deep SWS duration (min), REM duration (min), awake duration (min), sleep need (min), sleep consistency, nap count, nap duration, total duration (min), total energy burned (cal), weighted average heart rate (HR), weighted average HR zone 1, weighted average HR zone 2, weighted average HR zone 3, weighted average HR zone 4, weighted average HR zone 5, maximum HR (bpm), and weighted average HR zone 0. The following correlation matrix demonstrates how these variables relate to eachother within the final dataset.

Figure 6 shows the correlation matrix for the cleaned dataset.

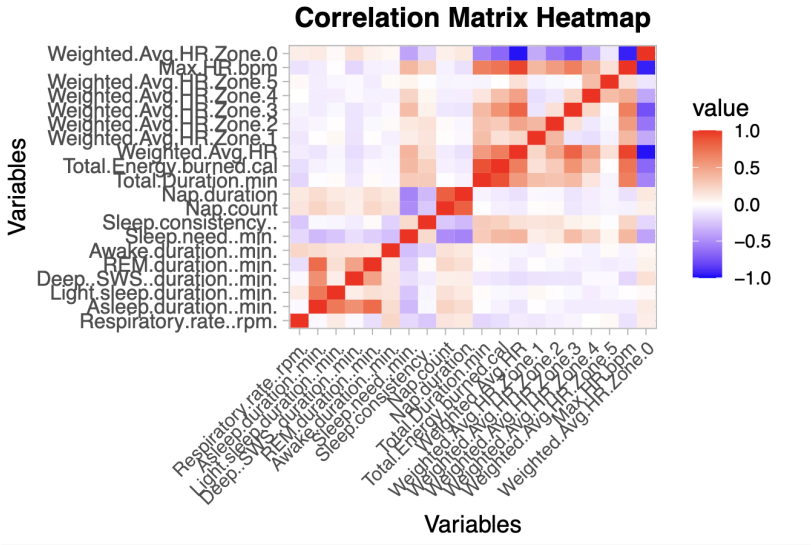

Figure 2: Correlation Matrix of the Cleaned Dataset.

2.6 EDA & Hypothesis

We initially hypothesized that **WHOOP's Activity Strain** (Activity Strain) would be the parameter with the biggest impact on the ratio of Deep (SWS) sleep to the total mount of sleep. However, after further thought we realized that the metric was likely a functin of data we already have. To examine this, we decided to use Activity Strain as the focus for our exploratory data analysis. In this EDA, we tested whether we needed Activity Strain to experiment with Deep SWS Sleep. The algorithm WHOOP uses to calculate Activity Strain is proprietary, but if other data could be used to strongly predict this measure then the metric would be collinear. In which case, it would be best to drop the metric entirely in an effort to avoid redundancy. In the next part of our EDA we attempted to model WHOOP's Activity strain via a linear regression.

To predict **WHOOP's Activity Strain**, we created a regression that had Activity Strain as our a dependent variable. Because our goal was to test the redundancy of including this metric, we used all variables that were not distance-related or time-related as our exploratory variables. This meant excluding GPS.enabled, Distance..meters., Altitude.gain..meters., Altitude.change..meters., Cycle.start.time, Cycle.end.time, Cycle.timezone, Workout.start.time, Workout.end.time, and Activity.name.

Linear Regression Model for Activity Strain

$$\text{Activity Strain} \sim \left(\begin{array}{r} \text{Heart Rate Avg (bpm)} + \text{Heart Rate Max (bpm)} + \text{Duration (min)} + \\ \text{Total Energy Burned (cal)} + \text{Weighted Avg HR} + \text{HR Zone 0} + \text{HR Zone 1} + \\ \text{HR Zone 2} + \text{HR Zone 3} + \text{HR Zone 4} + \text{HR Zone 5} \end{array} \right) \quad (1)$$

This high level of explanatory power suggested that the simplified model captures the essential predictors, such as heart rate data (max heart rate, average heart rate, and heart rate zones) and workout duration. We then decided to check the observe the correlation matrix for of the workouts dataset for further insight about how our predictors relate to one another.

Figure 3 shows the correlation matrix for the workouts dataset.

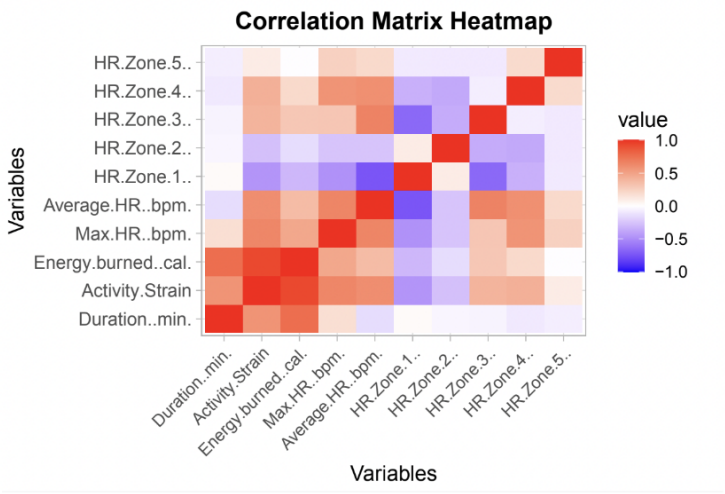

Figure 3: Correlation Matrix of the Workouts Dataset.

The correlation analysis of the workouts dataset was conducted using a threshold of 0.85 to identify highly correlated data. The resulting correlation graph revealed a strong correlation between Activity Strain and Energy Burned. To avoid issues of collinearity, the Activity Strain column was dropped, as Energy Burned serves as a reliable empirical measure of physical exertion. These changes were reflected in our final dataset.

3 Methods and Results

3.1 Methods for Inference Model

The baseline model was a generalized linear regression on the ratio of deep sleep to overall sleep across different days.

This regression relied on a comprehensive set of variables that included respiratory rate (`Respiratory.rate..rpm.`), total sleep duration (`Asleep.duration..min.`), deep sleep duration (`Deep..SWS..duration..min.`), nap count (`Nap.count`), nap duration (`Nap.duration`), total workout duration (`Total.Duration.min.`) total energy burned (`Total.Energy.burned.cal`), and heart rate data including the weighted average heart rate (`Weighted.Avg.HR`), weighted heart rate zone distributions (`Weighted.Avg.HR.Zone.0` through `Weighted.Avg.HR.Zone.5`) and maximum heart rate (`Max.HR.bpm`). There were no interaction terms. The aim for this diversity was to provide a comprehensive basis for predicting the ratio of deep sleep.

3.1.1 Assumptions for Baseline Model

There are a few things that must hold for linear regression to work here for inference. First, for linear regression we require that the response is continuous. The ratio we are trying to predict is continuous, but it is a fraction out of 1. But taking that into account would be out of the scope of this class. The second assumption is that the predictors are not highly correlated, and we can tell they are not from a correlation matrix from our EDA. Last, we first checked the standard assumptions: Existence, Linearity, Independence, and Homoscedasticity (ELIH). The results of these checks are visualized through residual diagnostics, including Residuals vs. Fitted plots, Q-Q plots of residuals, and histograms.

- **Linearity and Homoscedasticity:** The Residuals vs. Fitted plot (Figure 4) shows that the residuals are largely centered around zero with no clear non-linear pattern. However, there is slight heteroscedasticity, as evidenced by uneven spread near the center of the fitted values.

- **Normality of Residuals:** The Q-Q plot (Figure 4) indicates that the residuals follow an approximately normal distribution, with minor deviations in the tails caused by a few outliers.
- **Independence of Residuals:** The Residuals vs. Total Duration plot (Figure 5) demonstrates that residuals are randomly scattered, suggesting no relationship between the residuals and predictor values.
- **Residual Distribution:** The histogram of residuals (Figure 5) shows a roughly bell-shaped curve, supporting the assumption of normally distributed errors.

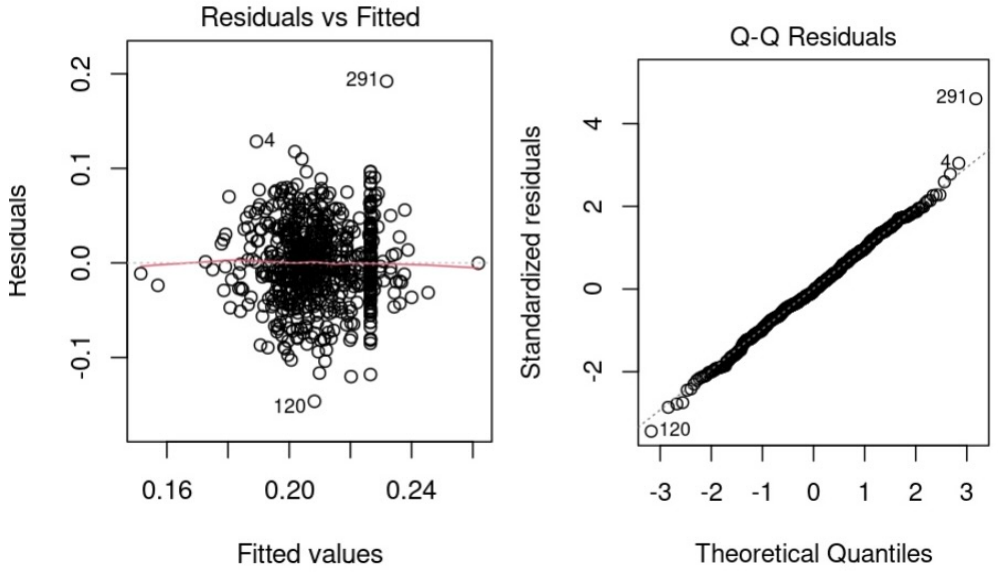

Figure 4: Residuals vs. Fitted (left) and Q-Q Plot of Residuals (right).

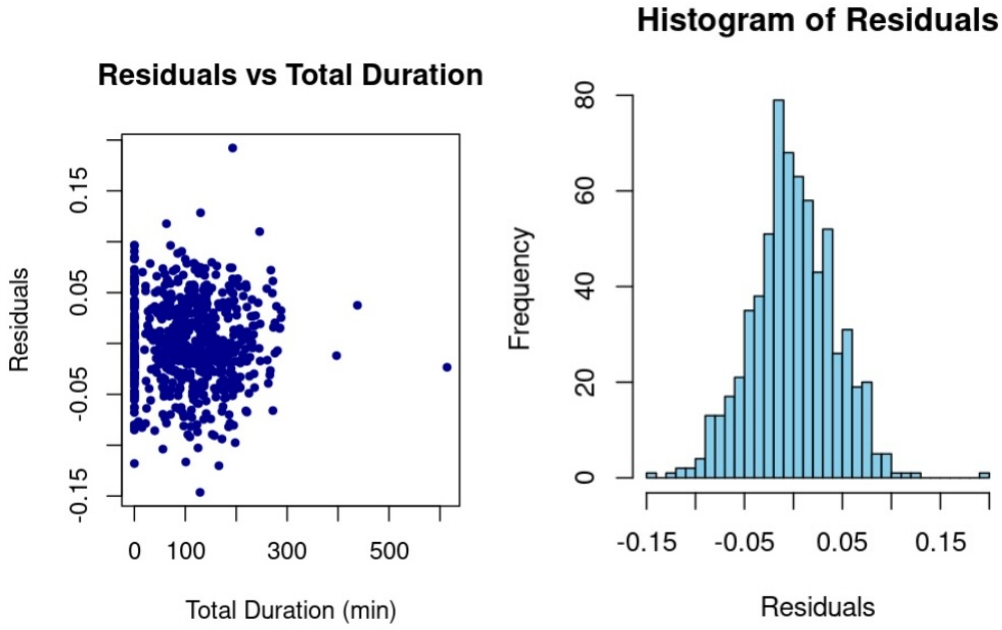

Figure 5: Residuals vs. Total Duration (left) and Histogram of Residuals (right).

The residual diagnostics confirm that the assumptions of linearity, independence, and normality are approximately satisfied. Minor deviations in homoscedasticity and the presence of a few outliers should be addressed in future models through transformations or robust regression techniques.

3.1.2 Results for more Sophisticated Models

We implemented one baseline model and two sophisticated models for inference in our analysis. The baseline model was a the most basic version of a linear regression designed to predict the ratio of SWS (Deep) sleep to overall sleep using the given parameters. The parameters of the model are Respiratory Rate (rpm), Light Sleep Duration (min), REM Duration (min), Awake Duration (min), Sleep Need (min), Sleep Consistency, Nap Count, Nap Duration, Total Duration (min), Total Energy Burned (cal), Weighted Avg HR, HR Zone 1, HR Zone 2, HR Zone 3, HR Zone 4, HR Zone 5, Max HR (bpm), and HR Zone 0. It had no interaction terms. The **adjusted R^2 value** of the baseline model was 0.347**.**

Our **first attempt** at a more sophisticated model included an interaction term with `Total.Energy.burned.cal`. This variable captures the total calories burned during the day, and is often reflects the total physiological load of the day. In the code file, this model is `model_rr`.The model uses all predictors in the dataset (like our baseline model) but builds on this model by including an interaction using the total energy burned during the day. Like the baseline model, the response variable is Deep..SWS..duration..min. / Asleep.duration..min., which represents the proportion of sleep spent in deep sleep. The **adjusted R^2** value of the interaction effects this model was 0.351, which is significantly higher than our baseline model (0.347).

By including the interactive term, the model is able to capture both the direct effect of total energy burned on the deep sleep ratio and its interactions with every other variable in the dataset. These interactions allow the model to evaluate how the relationship between each predictor and the deep sleep ratio changes based on the total energy burned. For instance, we theorized that the influence of variables such as average heart rate or nap duration on deep sleep may vary depending on the amount of energy expended during the day.

With our **second attempt** at a more sophisticated model, we developed a linear regression that predicts the deep sleep ratio by incorporating all main effects and two-way interactions between the predictors. In the code file, this model is `model_full`. While this approach captures more intricate relationships, it also introduces risks such as overfitting and reduced interpretability due to the interaction terms. In particular, we may have created high collinearity among predictors. The **adjusted R^2 value** for this model is 0.364.

Given the adjusted R^2 values and overall complexity, it makes sense that the more sophisticated models would be suitable candidates for backward stepwise selection. Backward step selection systematically simplifies the model by eliminating the insignificant interaction terms. To evaluate the quality of the models at each step of this process, we employed the Akaike Information Criterion (AIC), which assesses model performance relative to other potential models.

Originally when we ran these models, we had a higher adjusted R^2 value in the interaction model. Even though adjusted R^2 adjusts for the total number of parameters used, this outcome didn't seem likely. After this, we re-examined our code and made corrections. However, to ensure that we would conclude with the most optimal model we could find, we decided to conduct the stepwise selection process twice: once using each more-sophisticated model. The first more-sophisticated model yielded an AIC of **-4415.95**, which is notably lower than the AIC of **-4295.39** obtained for the second more-sophisticated model. Below is a representation of the models that resulted when backward selection was applied to each more-sophisticated model.

Backward Selction, First More-Sophisticated Model: $\frac{\textbf{Deep SWS Duration (min)}}{\textbf{Asleep Duration (min)}}$

$$\sim \text{Total Energy Burned (cal)} \times$$

$$\left(\begin{array}{r}\text{Respiratory Rate (rpm)} + \text{Light Sleep Duration (min)} + \text{REM Duration (min)} + \text{Awake Duration (min)} + \\ \text{Sleep Need (min)} + \text{Sleep Consistency} + \text{Nap Count} + \text{Nap Duration} + \\ \text{Total Duration (min)} + \text{Total Energy Burned (cal)} + \text{Weighted Avg HR} + \text{HR Zone 1} + \\ \text{HR Zone 2} + \text{HR Zone 3} + \text{HR Zone 4} + \text{HR Zone 5} + \text{Max HR (bpm)} + \text{HR Zone 0}\end{array}\right) \tag{2}$$

Backward Selection, Second More-Sophisticated Model: $\frac{\textbf{Deep SWS Duration (min)}}{\textbf{Asleep Duration (min)}}$

$$\sim \left(\begin{array}{r}\text{Respiratory Rate (rpm)} + \text{Light Sleep Duration (min)} + \text{REM Duration (min)} + \text{Awake Duration (min)} + \\ \text{Sleep Need (min)} + \text{Sleep Consistency} + \\ \text{Nap Count} + \text{Nap Duration} + \text{Total Duration (min)} + \\ \text{Total Energy Burned (cal)} + \text{Weighted Avg HR} + \text{HR Zone 1} + \\ \text{HR Zone 2} + \text{HR Zone 3} + \text{HR Zone 4} + \text{HR Zone 5} + \\ \text{Max HR (bpm)} + \text{HR Zone 0}\end{array}\right)^2 \tag{3}$$

Backward selection is a stepwise regression technique that begins with the most complex model. This means including all predictors. At each step, it evaluates the impact of removing one predictor using a metric such as AIC, and the least significant predictor is removed. This process is repeated until removing any further predictors would reduce the model's performance. Therefore, as a whole, this method helps to simplify models while retaining only the most significant predictors. Given this logic, it makes sense that by using this method, our second more-sophisticated model was able to reach the best adjusted R^2 so far.

Table 1: Adjusted R^2 Values for Different Models

Model	**Adjusted R^2**
Baseline Model	0.347
First Sophisticated Model	0.351
Second Sophisticated Model	0.364
Backward Stepwise (First Sophisticated Model)	0.363
Forward Stepwise (First Sophisticated Model)	0.353
Backward Stepwise (Second Sophisticated Model)	0.419
Forward Stepwise (Second Sophisticated Model)	0.367

Given this table, we came to the following conclusions. In both models, backward selection produced a higher adjusted R^2 value. Through step selection, the first more-sophisticated model improved from 0.351 to 0.363 and the second more-sophisticated model improved from 0.364 to 0.419. It should be noted that in both models, backward selection had a higher adjusted R^2 value than forward selection. The adjusted R^2 for the most optimal model was 0.419, which makes sense given that backward selection is designed to start from a highly complex model. This makes sense because backward selection is designed to start with the most complex model, and our second more-sophisticated model was more complex than any other model we have dealt with so far in this piece.

One potential hazard to be wary of is that interaction terms may exacerbate multicollinearity issues, especially if the predictors involved in the interaction are correlated. It is important to note that high multicollinearity can lead to inflated standard errors and less accurate coefficient estimates.

*Please see **Appendix** for the fitted/residuals plots and QQ plots for each of the odels referenced in this section.*

3.2 Inference on significant predictors

For our baseline model that explains a decent amount of the variance in the response, we also relay what we found from linear regression to be significant predictors at the 0.05-level.

Table 2: Baseline Model Outputs from `lm`.

Predictor	**Estimate**	**Std. Error**	**t Value**	**p-Value**
Intercept	-30.97	75.89	-0.408	0.6834
Respiratory Rate (rpm)	0.00001098	0.002646	0.004	0.9967
Light Sleep Duration (min)	-0.0005288	0.00003224	-16.399	$< 2 \times 10^{-16}$ ***
REM Duration (min)	-0.00006081	0.00003578	-1.700	0.0897$^{\cdot}$
Awake Duration (min)	0.0000865	0.00004466	1.937	0.0532$^{\cdot}$
Sleep Need (min)	0.00003521	0.00004005	0.879	0.3796
Sleep Consistency	-0.0002069	0.0001093	-1.893	0.0588$^{\cdot}$
Nap Count	-0.004514	0.007226	-0.625	0.5324
Nap Duration	0.00005424	0.0001443	0.376	0.7071
Total Duration (min)	0.00008922	0.00007057	1.264	0.2066
Total Energy Burned (cal)	-0.00001293	0.00001236	-1.046	0.2958
Weighted Avg HR	0.000444	0.0004750	0.935	0.3503
Weighted Avg HR Zone 1	0.3125	0.7590	0.412	0.6807
Weighted Avg HR Zone 2	0.3126	0.7590	0.412	0.6805
Weighted Avg HR Zone 3	0.3126	0.7590	0.412	0.6806
Weighted Avg HR Zone 4	0.3128	0.7590	0.412	0.6804
Weighted Avg HR Zone 5	0.3130	0.7590	0.412	0.6802
Max HR (bpm)	-0.0003551	0.0001559	-2.278	0.0230*
Weighted Avg HR Zone 0	0.3127	0.7590	0.412	0.6805

The duration of light sleep, the number of hours awake, the WHOOP-calculated sleep needed, and the duration of naps during the day (0 when there are no naps) were all statistically significant predictors. Perhaps unsurprisingly, in a basic model with no interaction terms, the predictors that were most important to predict the ratio of deep sleep to total sleep were all sleep-related, not exercise-related. Intuitively, it makes sense that one could explain the proportion of deep sleep by using the light sleep duration as an inverse correlative variable.

Likewise, to calculate their metric for sleep-need, WHOOP uses a combination of the person's baseline sleep need (generally 7.6 hours), physical exertion through the day (strain), previous sleep debt, and naps. Additionally, it would also make sense for that person to have more time spent in Deep (SWS) sleep. [4] [3]

Our best model had too many predictors to fully analyze its output in this piece, however, we included a table of significant predictors and their respective p-values. While there were many more interaction terms, there were also more variables were statistically significant in this model that related to exercise. To give a couple examples, this included the interaction between the total energy burnt and the weighted average heartrate. The complete table of the significant predictors from our best model (and their varying levels of significance) is below.

Table 3: Significant Coefficients and Levels of Significance

Predictor	Estimate	Std. Error	p-value	Significance Level
Respiratory.rate..rpm.:Awake.duration..min.	-2.459e-04	1.083e-04	0.023642	*
Light.sleep.duration..min.:Sleep.consistency..	-8.457e-06	2.910e-06	0.003823	**
Total.Energy.burned.cal:Weighted.Avg.HR	1.987e-05	5.936e-06	0.000881	***
Weighted.Avg.HR.Zone.1:Weighted.Avg.HR.Zone.4	-2.408e-04	1.103e-04	0.029573	*
Weighted.Avg.HR.Zone.2:Weighted.Avg.HR.Zone.3	-6.505e-05	2.810e-05	0.021019	*
Weighted.Avg.HR.Zone.3:Weighted.Avg.HR.Zone.0	-1.624e-04	7.313e-05	0.026864	*

3.3 Methods for Prediction Models

Our approach to try and do prediction will use a different response variable for what we are trying to capture: *a good nights' sleep*, which is still related to the ratio of deep-sleep to total sleep. We will instead classify days into good-sleeps and bad-sleeps. It seemed like the median of days had a deep-sleep-ratio of 0.211. so we added a binary response variable that is 1 for all sleeps with this variable above 0.211 (n=235), and 0 for sleeps with the variable below 0.5 (n=236). Using a binary response will allow us to evaluate the results of our prediction more meaningfully, as a mean-squared error of prediction on a response that is a number between 0 and 1 is not sensible. Binary prediction is still quite useful as the core question we are trying to get at is whether we can predict whether someone will have good quality sleep based on their exercise patterns.

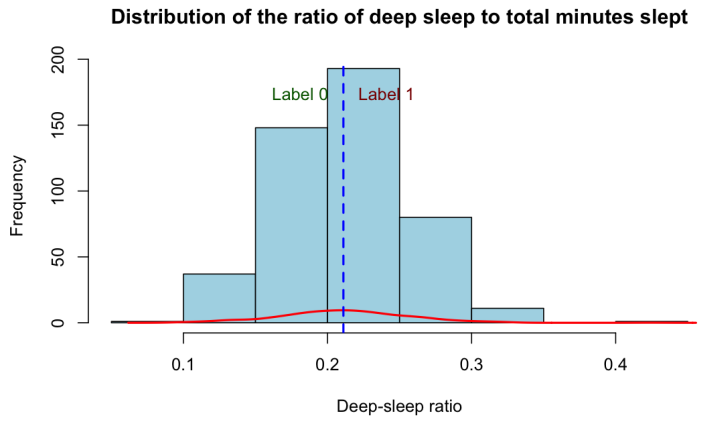

Figure 6: Histogram for the distribution of deep sleep.

3.3.1 Data Splitting & Model Evaluation

Across the observations (days), We used a seed of 139 to maintain reproducibility across the different predictive models we tried. Across all days, we randomly split 80 percent of our data into the training set and 20 percent into the testing set. Our dataset is reasonably balanced, so accuracy should be a good metric of our binary predictors' performance, but we have also included the metrics of precision, recall, and F1-score, the harmonic mean of precision and recall.

3.3.2 Baseline Predictive Linear Model

We use a simple logistic regression model from the `glm` package with the binomial family. If the classifier predicts a probability of greater than 0.5 for a sleep on that day, based on the characteristics of the exercise done that day, of being classified as a good sleep, then the model will classify it as a good-sleep day. The

assumptions of logistic regression hold here: we are predicting a binary variable, we have a large enough sample size (n=471), and we have little multicolinearity among our predictors.

3.3.3 Additions to logistic regression

We experimented with the following models with the following settings for binary prediction:

1. LASSO regression: to shrink coefficients with less predictive power, to prevent overfitting. We tune the regularization parameter λ using cross-validation.
2. k-nearest neighbors: predicting the label based on the majority class of the k nearest data points seen in training. We tune the hyperparameter k by cross-validating over 10 values of k.
3. Random Forest: Used to test efficacy of tree-based against the linear methods we focus on in this class. We used cross-validation to set `mtry,` the number of predictors selected for consideration at each split.
4. Gradient-Boosted Trees: Also as a point of comparison against linear models, we use cross-validation to select nrounds and the maximum depth of each tree in the ensemble.
5. Support Vector Machines. We tune cost of misclassification and RBF kernel width by cross-validation.

3.3.4 Prediction Results

Table 4: Model Performance Metrics on the Test Set

Model	Precision	Accuracy	Recall	F1
Logistic Regression (Baseline)	0.573	0.639	0.729	0.642
Logistic Regression w/ LASSO	0.581	0.647	0.729	0.647
K-Nearest-Neighbors	0.551	0.617	0.729	0.628
Random Forest	0.556	0.617	0.678	0.611
Gradient-Boosted Trees	0.589	0.654	0.729	0.652
Support Vector Machines	0.577	0.647	0.763	0.657

We present a table of the performance on the hold-out test set of accuracy, precision, recall and F1-score, for the 6 models we fit. As we can tell, the models all have 60-plus percent accuracy and F-1 score. Interestingly, the tree-based methods do not present a significant advantage over the linear models here. All models perform in the same ballpark, with gradient-boosting providing a small edge in terms of accuracy and F-1. All models perform slightly better than the naive baseline of 50% accuracy.

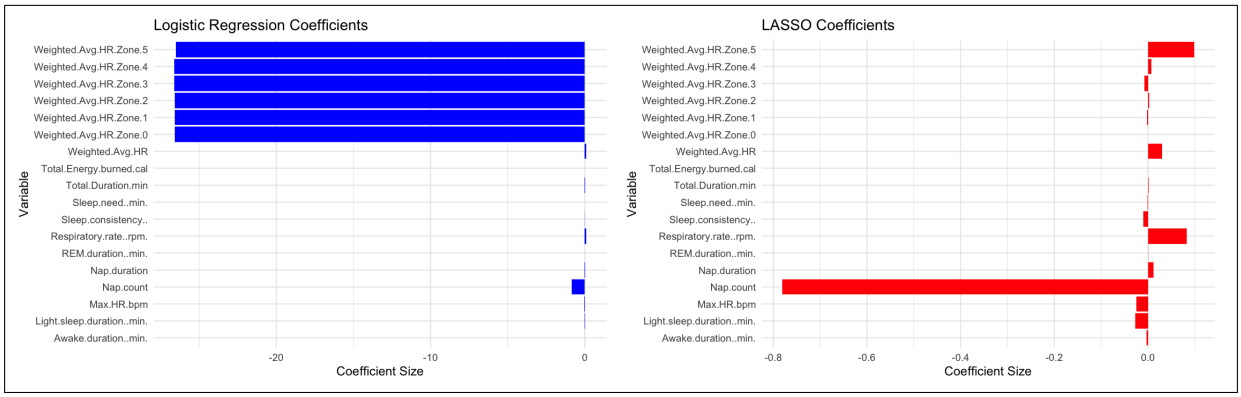

Figure 7: Comparison of Absolute Coefficient Sizes: Logistic vs LASSO Regression.

We have plotted the relative sizes of the coefficients for each model separately in the below two charts. We see that LASSO shrunk the variables `Weighted.Avg.HR.Zone.0`, and `Total Energy Burned` to zero.

Interestingly, LASSO assigns a high relative importance to the number of naps taken during the day to predict the quality of sleep.

4 Conclusion and Further Work

We conclude that some of the quality of an individual's sleep can be measured from their exercise patterns, especially from the perspective of binary prediction on whether someone slept well. We did find certain statistically significant associations though between exercise patterns and one's sleep. One area of further research would be to find a great metric for the quality of one's sleep. Another interesting experiment could be to include more binary data, based on the type of exercise done during the day, for example, to see if that affects the relationship between predictor and response.

Ultimately, a number of improvements could be made in the future that would continue our research. Firstly, collecting predictive variables closer to the time of one's sleep may ultimately prove more useful in pure prediction of sleep quality. Furthermore, the relatively small size of our dataset in terms of observations and the fact that we are limited to the data of one individual means that our prediction method will likely not generalize well to the population, and is perhaps not even robust to the sleep Eva is likely to have one year from now. Also, we should properly limit ourselves to strictly data from exercise-related activity.

Still though, our work suggests that some insights can be gained for improving sleep quality based on what is done during the day. We saw for example, surprisingly, that one's overall weighted average heart-rate zone is statistically significant and positively correlated with the ratio of deep sleep to total sleep.

References

[1] Will Ahmed. Whoop approved for in-game use in major league baseball. *WHOOP*, March 2017.

[2] CPerlman. Whoop: Using data to target the elite athlete. *Digital Innovation and Transformation*, 2024. Accessed 17 Dec. 2024.

[3] WHOOP. Understanding sleep debt: Impact on performance and recovery. *WHOOP Locker*. Accessed December 21, 2024.

[4] WHOOP. Whoop sleep: Understanding your sleep metrics. *WHOOP Support*. Accessed December 21, 2024.

[5] WHOOP. What is restorative sleep? *WHOOP*, October 2023.

5 Appendix: Parameter Distributions & Model Plots

A GitHub page with all of our experiments and data for replication can be found at our GitHub Repository

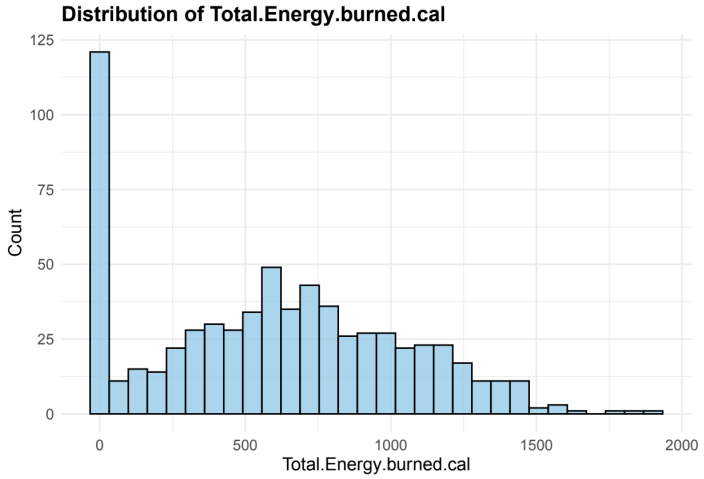

Figure 8: Distribution of Total.Energy.burned.cal

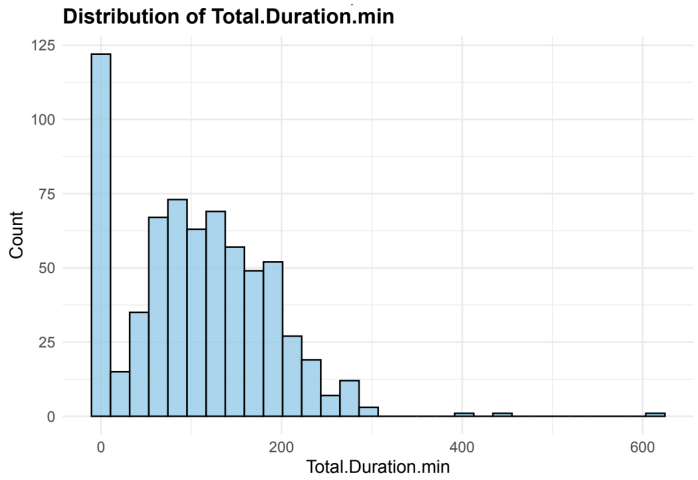

Figure 9: Distribution of Total.Duration.min

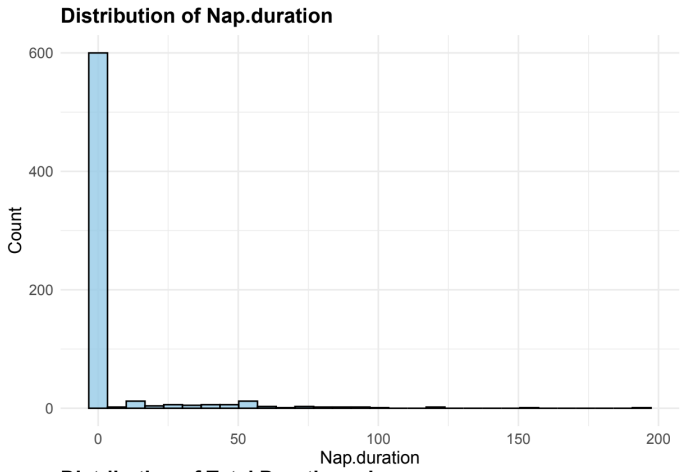

Figure 10: Distribution of Nap.duration

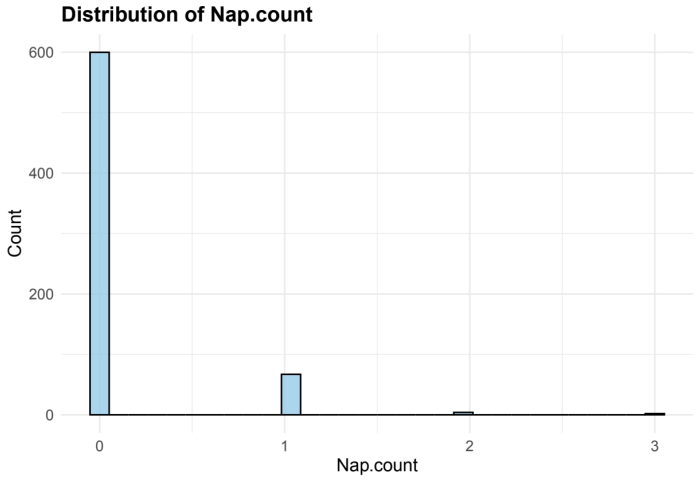

Figure 11: Distribution of Nap.count

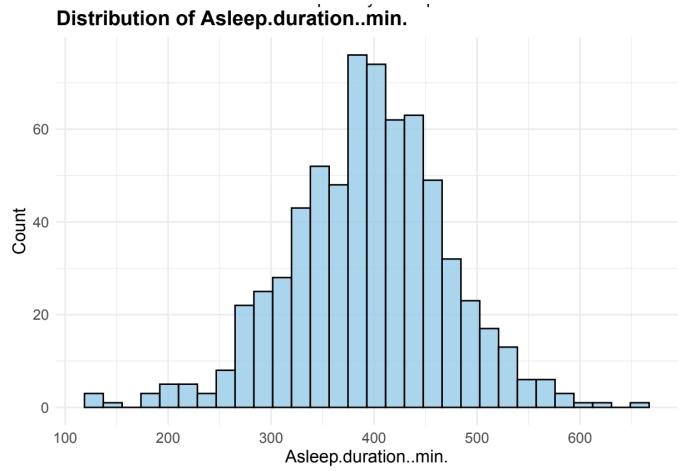

Figure 12: Distribution of Asleep.duration..min.

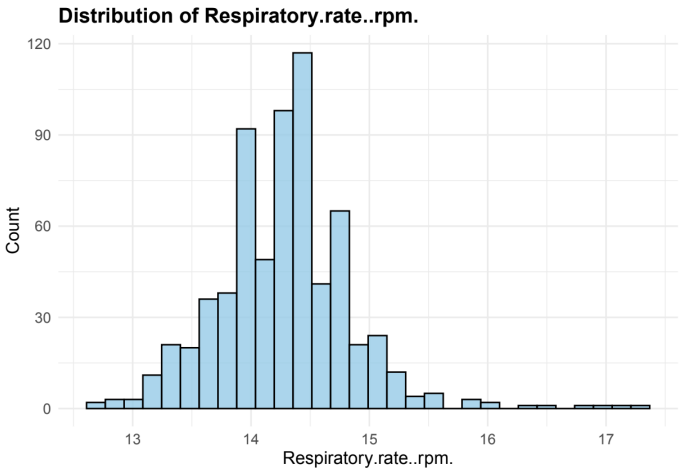

Figure 13: Distribution of Respiratory.rate..rpm.

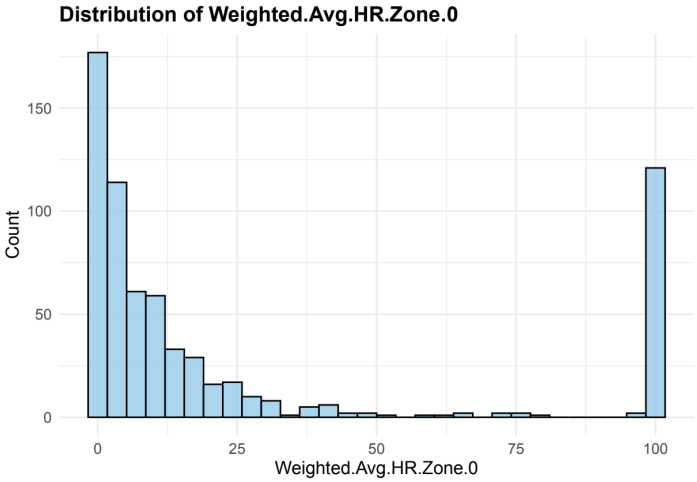

Figure 14: Distribution of Weighted.Avg.HR.Zone.0

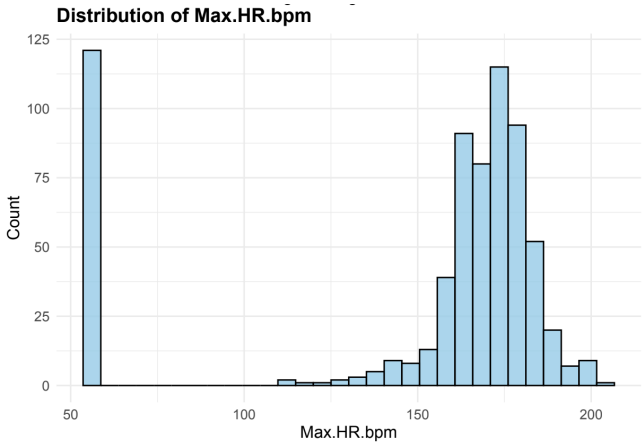

Figure 15: Distribution of Max.HR.bpm

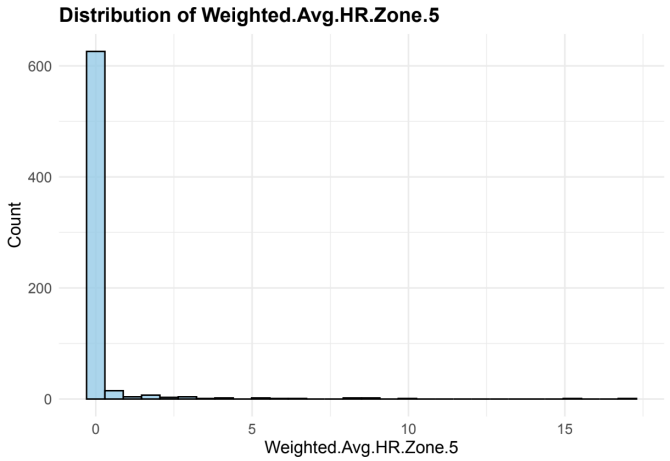

Figure 16: Distribution of Weighted.Avg.HR.Zone.5

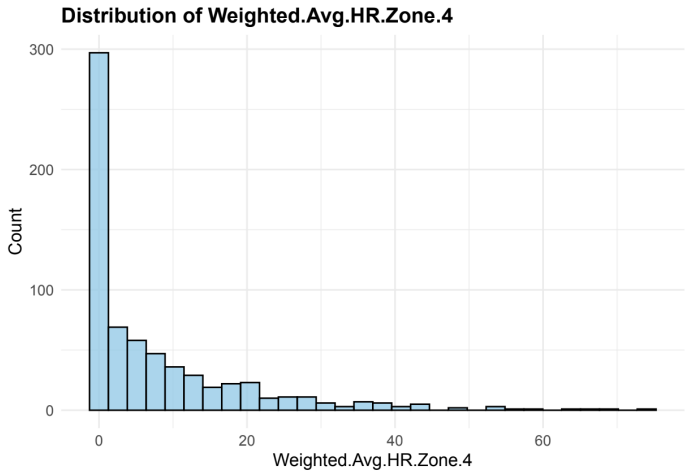

Figure 17: Distribution of Weighted.Avg.HR.Zone.4

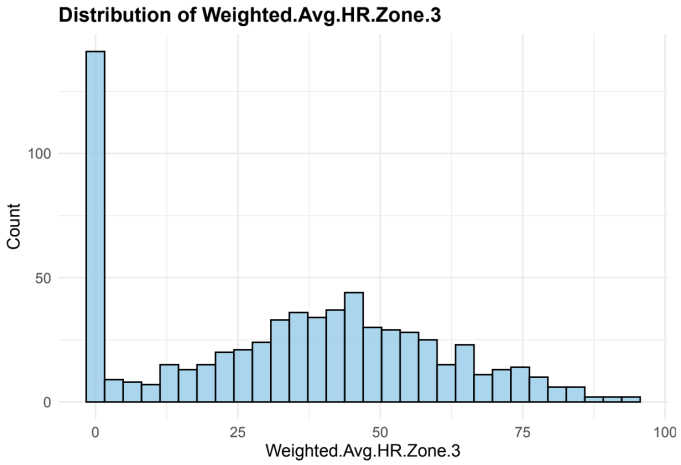

Figure 18: Distribution of Weighted.Avg.HR.Zone.3

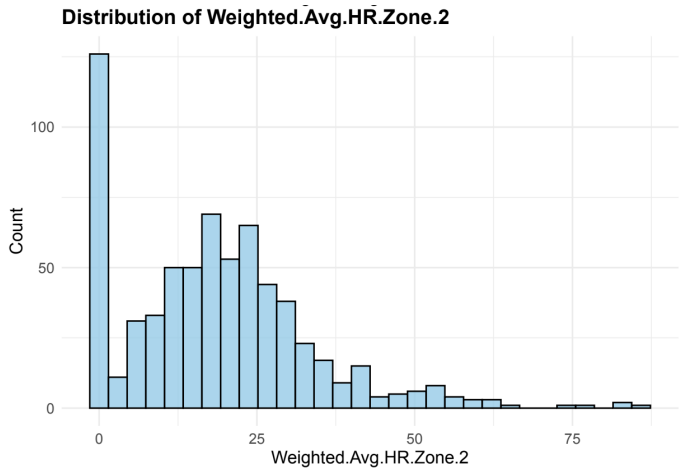

Figure 19: Distribution of Weighted.Avg.HR.Zone.2

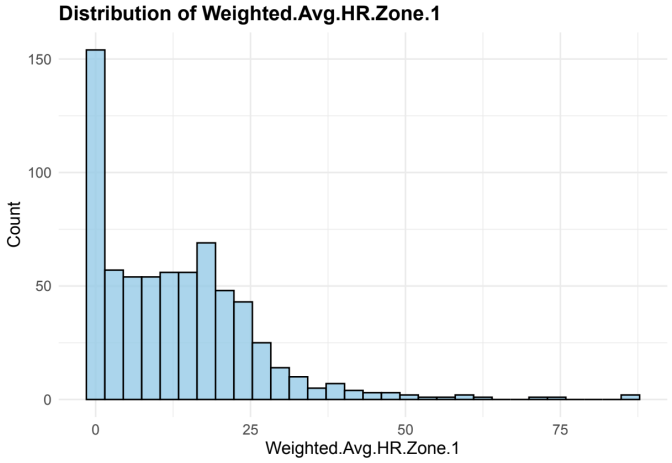

Figure 20: Distribution of Weighted.Avg.HR.Zone.1

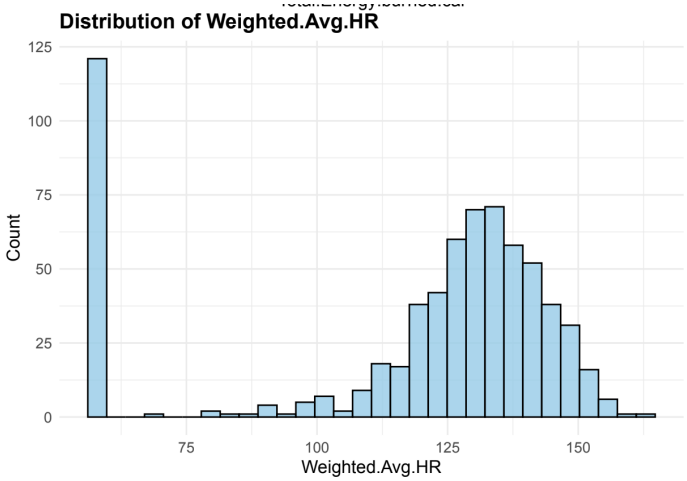

Figure 21: Distribution of Weighted.Avg.HR

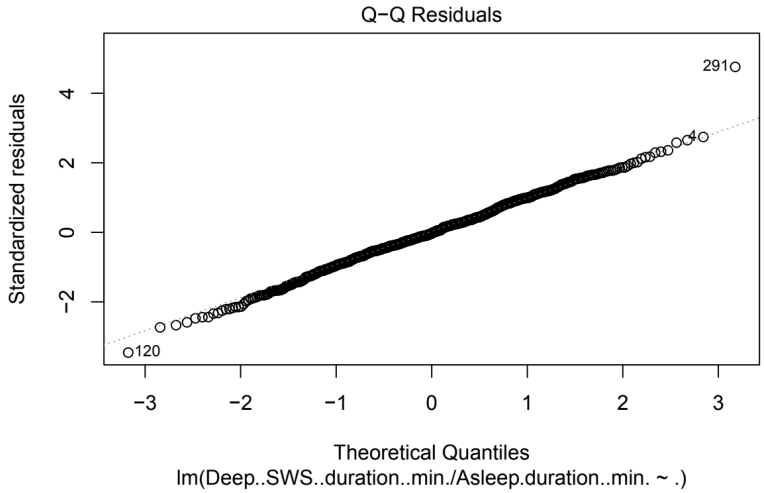

Figure 22: General Model: Residuals vs Fitted

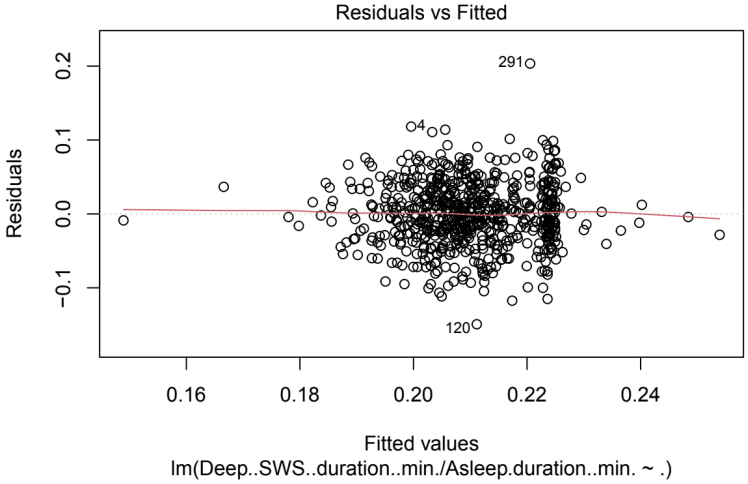

Figure 23: General Model: Q-Q Residuals

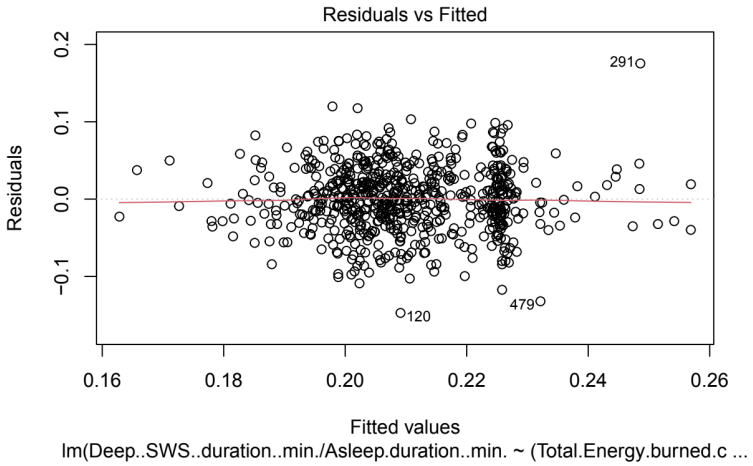

Figure 24: Interactions Model: Residuals vs Fitted

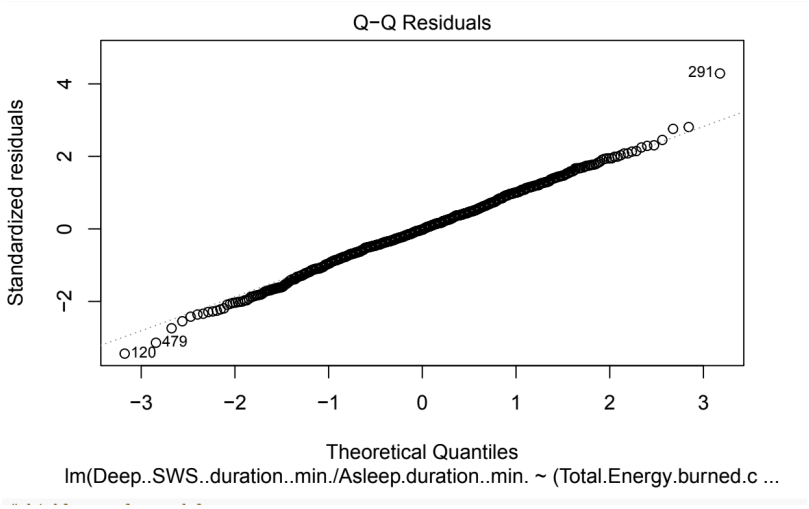

Figure 25: Interactions Model: Q-Q Residuals

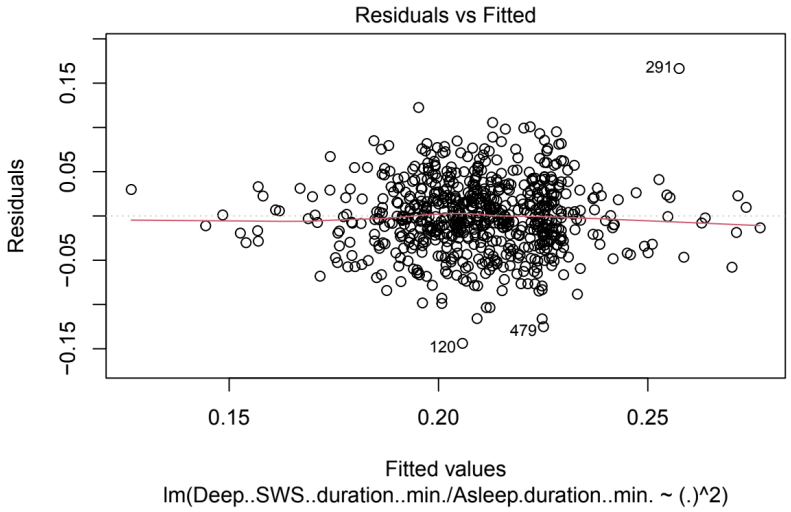

Figure 26: Full Model: Residuals vs Fitted

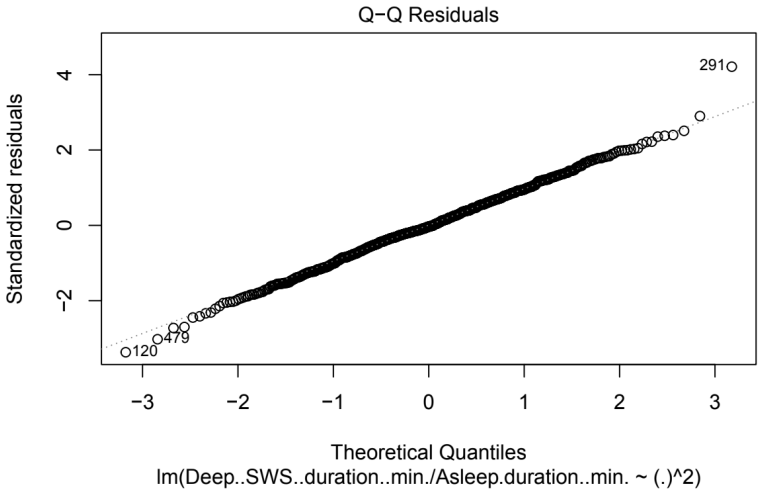

Figure 27: Full Model: Q-Q Residuals

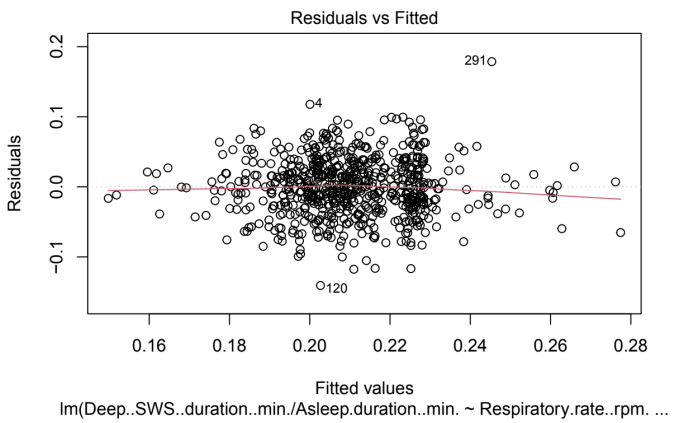

Figure 28: Model Selection (Backward): Residuals vs Fitted

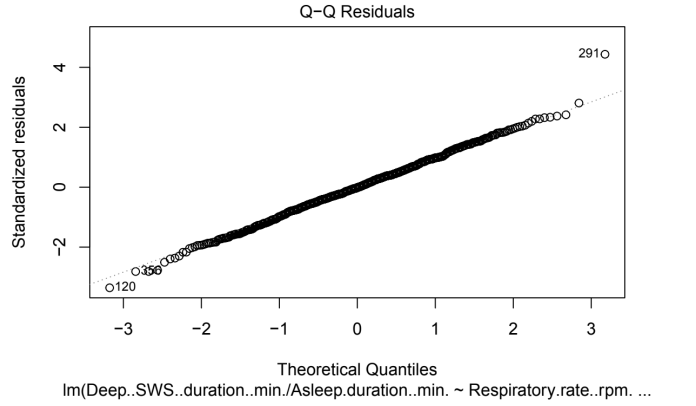

Figure 29: Model Selection (Backward): Q-Q Residuals

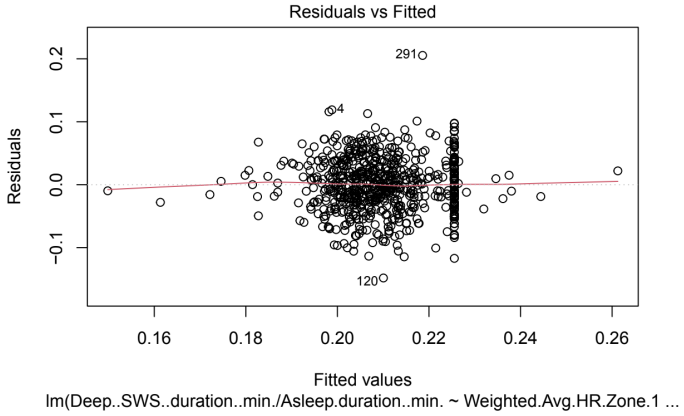

Figure 30: Model Selection (Forward): Residuals vs Fitted

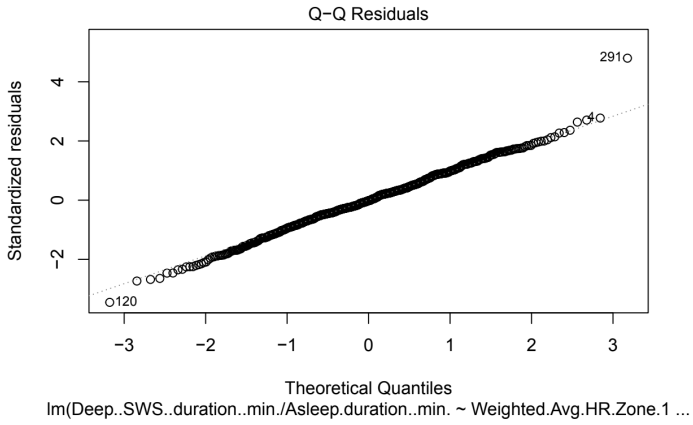

Figure 31: Model Selection (Forward): Q-Q Residuals

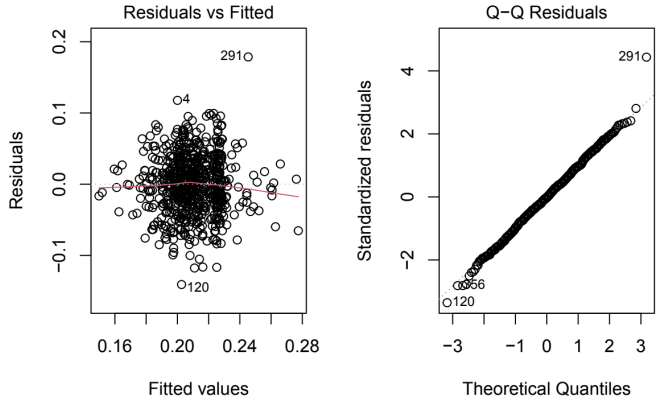

Figure 32: Prediction: Residuals vs Total Duration and Histogram of Residuals

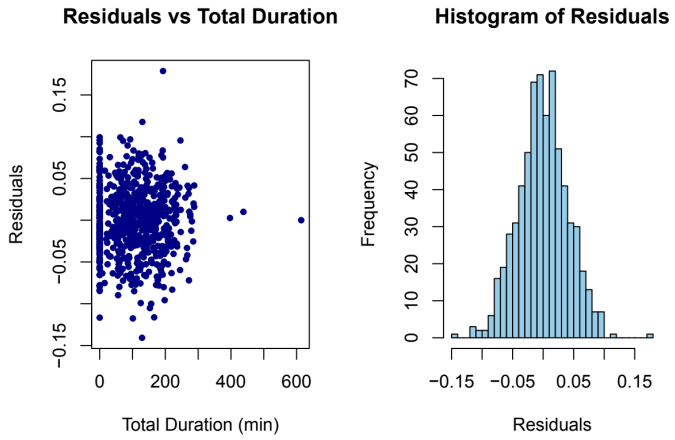

Figure 33: Prediction: Residuals vs Fitted Values and Q-Q Plot of Residuals