Problem 1: TriangleWell
# Logistic Regression and Analysis

## Contents

1. **Summary**  
2. **Introduction**  
   2.1 Objectives  
   2.2 Assumptions  
3. **Model Creation**  
   3.1 Data Preparation  
   3.2 Variables  
   3.3 Logistic Regression  
   3.4 Smoothing  
   3.5 Validation  
4. **Analysis**  
5. **Conclusion**  
   5.1 Strengths and Weaknesses  
   5.2 Future Work  
6. **Letter to President and Wellness Center**  
7. **Appendix**

---

## 1 Summary  

---

## 2 Introduction  

2.1 Objectives  

---

## 3 Model Creation  

3.1 Data Preparation  
3.2 Variables  
3.3 Logistic Regression  
3.4 Smoothing  
3.5 Validation  

---

## 4 Analysis  

---

## 5 Conclusion  

5.1 Strengths and Weaknesses  
5.2 Future Work  

---

## 6 Letter to President and Wellness Center  

---

## 7 Appendix  

---
1 Summary

We develop a set of logistic regression models to estimate sleep patterns of college students based on data from a pedometer recorded in 15-minute intervals. We further refine the logistic regression models with a smoothing algorithm to produce reasonable sleep patterns on a student-by-student basis while preserving overall model accuracy: on validation data, our model achieves an average of 84% classification accuracy, ranging between 59% and 98%. Applying the model to the prediction dataset, we determine that sleep time exhibits a slight negative correlation \( r = -0.224 \) with activity levels and that sleep time remains fairly constant throughout the week. Consequently, we recommend that college administrators and student health departments emphasize the importance of consistent, high-quality sleep, as opposed to inconsistent and/or excessive but low-quality sleep, while also highlighting the role of physical activity in promoting healthy sleep habits.

2 Introduction

2.1 Objectives

With an increased concern for student health on campus, many colleges have been directing attention to student health patterns and encouraging healthy habits. One such example involves the use of electronic devices to monitor physical activity, through the counting of steps. The TriangleWell developer committee has approached us to provide insight into student sleep habits, predicted from the number and timing of students’ steps. Our team seeks to develop mathematical and statistical models to take students’ step counts and infer when they are sleeping. Once this model has been created, we can provide TriangleWell with information about sleep habits and activity levels. We can look for associations between total sleep time and the number of steps taken the previous day and between bedtimes and the number of steps taken the next day. With this information, TriangleWell and universities can collaboratively work to promote both physical fitness and sleep among students.

2.2 Assumptions

Our model makes a number of assumptions:

1. The number of steps a student takes is independent of the number of steps other students take.

2. In order to ensure independence between observations from the same student, we can take the average of the steps of a given student, at a given time, and find the mode of the sleep label, for a given student, at a given time. Use of these summary statistics is representative of the student’s activity levels at that time of the day.

3. We designated the first 80 observations in the "Steps + Sleep" Google Sheet as the training dataset. The proportion of students asleep at a given time throughout the day in the training dataset is representative of the "Steps" dataset.

4. There is a linear relationship between the natural logarithm of the odds that a student is asleep.

5. Steps taken at each individual time interval across all days in October. This allows us to model the data with a logistic regression model.
6. If, between 11 PM and 6:30 AM, an "awake" observation at a particular 15 minute interval is surrounded, 15 minutes prior and afterwards, by an "asleep" observation, that student was actually asleep during that interval, and the sleep label was inaccurate.

7. If, between 6:30 AM and 11 PM, an "asleep" observation at a particular 15 minute interval is surrounded, 15 minutes prior and afterwards, by an "awake" observation, that student was actually awake during that interval, and the sleep label was inaccurate.

8. A student is said to wake up at the first time following 4 AM such that the model labels them as awake.

9. With steps only it impossible to differentiate between a period of inactivity (such as time spent seated in class or working in the library) from an afternoon nap. Therefore, we will assume students do not fall asleep before 5:30 PM, even if they record 0 steps. If a 0 step increment occurs before 5:30 PM, and after a student has woken up, they will be marked as awake. (Assumptions 6 through 8 are the basis of our "smoothing" algorithm we use to correct for the granularity of our model; their purpose and effect will be discussed after the model itself is outlined.)

10. When a student in the "Steps" dataset moves exactly 2500 steps in a 15-minute period, this is indicative of a device error, and this value is replaced with a 0 steps.

3 Model Creation

We began the process of model creation with an inspection of the "Steps + Sleep" dataset. The purpose of our model is to input information about the number of steps a student takes at a particular time of the day. Then, using the time of day and number of steps, the model would provide a prediction of whether or not the student was asleep at that time.

3.1 Data Preparation

The data were prepared for the modeling process by breaking the data into individual observations, so that there is one observation for each student on each day at each 15-minute interval, rather than one observation per student per day. Thus, for example, a single observation in the prepared data would specify Student ID, date, time, steps, and sleep status, and the prepared dataset contains a total of 100 × 31 × 96 observations.

For the purposes of training and evaluating the model, we separated the "Steps + Sleep" dataset into two components; all observations of the first 80 students (IDs 0-79) were put in the training dataset, while all remaining observations (IDs 80-99) were put in the validation dataset.

While we did observe a surprising number of observations recording exactly 2500 steps here, we decided not to replace these numbers as students were always marked as awake during intervals with 2500 steps recorded, meaning that these (likely erroneous) observations were unlikely to particularly affect the model (since any reasonable model will already mark anyone with such a high number of steps as awake).
3.2 Variables

Our model does not use many variables, as the goal is to produce estimates of sleep patterns based only on step data. As such, only the following variables are considered and used in the model. One observation in each dataset contains one value for each of the following variables:

1. **Student ID** gives a unique identifier for each student. Student IDs from 0 to 99 have associated sleep data and are used in the training and validation datasets; Student IDs from 100 to 499 do not have associated sleep data and are used solely in the prediction dataset (in the Analysis section of this report).

2. **Date** indicates the date on which the observation was taken—between October 1 and October 31, 2021.

3. **Time** records the time of day during which the observation was taken. Data is recorded in 15-minute intervals and the value of the time variable indicates the beginning of the 15-minute window corresponding to the observation.

4. **Steps** records the number of steps taken by the student during the given time period on the given day.

5. **Asleep** is a binary variable encoding whether the student reported being awake or asleep during the given time interval on the given day. This variable is only available for Student IDs 0-99 in the training and validation datasets.

6. **Prediction** records the model’s prediction of whether the student was awake or asleep during the given time period on the given day, based solely on the number of steps taken by the student and the time of day.

3.3 Logistic Regression

Logistic regression uses a logistic function to model a binary dependent variable, in our situation, this is the sleep label. In other words, logistic regression models the log-odds of the response variable as a linear function of the independent predictor variables. Hence, logistic regression assumes a linear relationship between the number of steps a student takes and the log-odds of the probability that a student is asleep in a given 15-minute interval. This assumption is not wholly unreasonable, although it is not necessarily guaranteed to be true.

We produced 96 distinct logistic regression models: one for each 15-minute interval. We found this to yield more accurate results than producing a single logistic regression model that used time as a predictor (either as a numeric or categorical variable), since such models usually ended up ignoring steps entirely as time is simply a better predictor. This is, however, also a weakness of the model, since it means that the model requires additional smoothing (discussed later) to make the predictions make sense on a student-by-student level. In other words, the model *a priori* does not produce particularly consistent results for a single student over the course of a day, since the predictions are produced using 96 distinct logistic regression models.

To create the 96 models, we separated the training data into 96 datasets, one for each 15-minute interval. Each of these datasets required further processing because logistic regression also assumes...
that the observations in the training data are independent. This assumption poses a challenge here, since the training data was certainly not independent—multiple samples from a single individual over the course of a month will naturally exhibit dependence. To mitigate this, we aggregated all 31 observations from each student within each time interval into a single observation. For instance, all observations of the student with ID 0 made at midnight were aggregated into a single observation in the midnight training dataset. The aggregation involved taking the mean of the steps taken and the mode of the sleep status across the 31 observations of the student at the time in question. As the aggregated observations now come from different students, it is reasonable to assume that they satisfy the independence requirement of logistic regression.

Logistic regression was then performed on each of the 96 training datasets. Each logistic model predicts the log-odds of a student being asleep based on their steps in the fixed 15-minute interval. The probability of a student being asleep was computed from the log-odds, and initially we assigned labels based on this probability, with the breakpoint set at 0.5: a probability less than 0.5 led to a label of 0 (awake) while a probability greater than 0.5 led to a label of 1 (asleep). Below is one of the 96 logistic models created from the 8:30 time interval. Larger number of steps correspond to a lower probability of being asleep.

![Logistic Regression at 8:30 AM](image)

**Figure 1: Example Logistic Regression at 8:30 AM**

Not all logistic models were well-behaved. Two typically failed to converge, and a number (particularly in the mid-late afternoon) essentially ignored steps and predicted that everyone would be awake. While this is not ideal behavior, we interpret it as a signal that steps simply are not a good predictor of sleep status at those times (at least, not as a linear predictor for log-odds). Indeed, the statistical significance ($p$-values) of the regression coefficients for those afternoon models are exceptionally high (some even reaching 1.00), indicating that, in those time intervals, the training data does not provide evidence that a correlation exists between steps and log-odds of being asleep.
3.4 Smoothing

Upon an inspection of the model predictions, we discovered that the model often produced "rough" or unrealistic sleep patterns. In particular, the models often had students alternating between asleep and awake for subsequent 15-minute intervals even while the student was actually asleep for all intervals in question. Similarly, the model also often indicated that students were asleep for isolated 15-minute intervals during the day, which is exceptionally unlikely (and, in any case, virtually impossible to distinguish from simply sitting for 15 minutes based on step data alone). Therefore, we decided to apply a smoothing algorithm to make the results more consistent with natural sleep patterns.

The smoothing algorithm switched the predicted sleep status of a student in a given 15-minute interval under certain specific conditions. A label of 0 (awake) was switched to a label of 1 (asleep) between 11pm and 6:30am if the label of 0 was both preceded and followed by a predicted label of 1 (in the adjacent 15-minute intervals). Similarly, a label of 1 (asleep) was switched to a label of 0 (awake) between 6:30am and 11pm when it was surrounded on either side by labels of 0 (awake). This assumed that students do not engage in power naps, which is a mostly reasonable and a necessary assumption because steps alone cannot provide sufficient evidence to conclude that a student takes an isolated 15-minute nap during the day. The smoothed labels were stored in a new variable, and only the original labels were used in the smoothing process; this ensured that smoothing did not cascade throughout the data, and swap far more labels than intended. The problem statement suggested there may be device noise and user error, supporting our claim that the data needed to be smoothed and edited to properly train and evaluate our model.

Using assumptions 8 and 9, we also ran a second smoothing algorithm to remove sporadic daytime naps. This algorithm found the first time after 4 AM that a student was predicted as awake, and assumed that this was the time at which they woke up for the day. All subsequent periods during that day occurring before 5:30 PM that were labeled as asleep by the logistic models were changed to being labeled as awake, in light of the insufficiency of step data alone to determine daytime napping habits. This more drastic form of smoothing also helped make the data more reasonable and more accurate by decreasing the extent to which the algorithm overestimated daily sleep.

While we did also consider attempting to correct unreasonable step counts in the training data in light of the problem’s indication that such error was likely, we decided that doing so would harm the model’s ability to predict sleep patterns in real-world data, which is likely to contain similar inaccuracies in the step counts.

3.5 Validation

The model was validated using the remaining 20% of the "Steps + Sleep" dataset. We first computed the predictive accuracy of the model, both with and without smoothing, at each of the 96 individual time intervals. We found that the model without smoothing ranged between 52 and 98 percent accuracy at 9am and 7pm, respectively, with an average accuracy of 82%. The smoothed model ranged between 59 and 98 percent accuracy at 11:30pm and 7pm, respectively, with an average accuracy of 84%. (See Figure 2.)

While the overall accuracy is 84%, this value greatly differs by time interval. The model is most accurate between 12:15 PM and 7:45 PM. In this time frame, the model is 97% accurate. During these time periods, the coefficient on the steps variable is extremely small, indicating that the model does not rely much on the number of steps taken. Instead the model is accurately predicting that students are
3.5 Validation

![Smoothed vs. Unsmoothed Model Performance](image)

**Figure 2: Model performance**

Our model is least accurate between 7:15 AM and 10:45 AM and between 10:30 PM and 1:00 AM. This is to be expected, as during these times students are waking up and falling asleep. This puts more emphasis on the steps variable as a predictor of whether or not the student is awake, as the time period itself is not as good a predictor of whether or not a student is awake. During both of these times, the accuracy of the model is, on average, 65%. However, the average coefficient on the steps term in our model is -0.099 and -0.0479, which is relatively realistic. This shows that the model takes the number of steps a student records into account when making its prediction. All models for time periods between 7:15 AM and 10:45 AM and for time periods between 10:30 PM and 1:00 AM have coefficients on the steps variable that are significant at an alpha level of 0.05, emphasizing that our model does make use of the steps variable at these times.

Another indicator that our model is fairly accurate is the similarity between the amount of sleep that our model predicts and the true amount of sleep that was reported in the final RPE of the "Steps + Sleep" observations, as shown in Figure 3. Our model tends to over-predict sleep, by an average of 0.854 hours per day.

In the course of validation, we also manually examined patterns among specific students. Our model is not equipped to handle students who wake up early, but do not take many steps until later in the day. For example, student 80 wakes up on October 1st at around 4:45 AM, however only begins to take consistently large numbers of steps beginning at 6:45 AM. Our model has no method to know if a student is awake if they do not take many steps.

While the smoothing appears to slightly decrease the model accuracy in the early morning hours, it nonetheless improves the model’s predictions by making the results, considered on a student-by-student basis, seem reasonable. In particular, it removes discrepancies where an individual would appear to
oscillate between asleep and awake over the course of the night when in fact they were really asleep. So the decrease in overall accuracy in these time periods is offset by the more reasonable predictions.

4 Analysis

We now proceed to analyze patterns in the non-labeled dataset by applying the models developed thus far. We first apply all of the same data transformations described in the Data Preparation section. Now, we can predict whether or not a student is asleep using our model. We were only able to predict sleep labels for individuals labeled between 100 and 150 due to computational constraints—in particular, the smoothing algorithms were exceptionally slow, with runtimes of upwards of 30 seconds per student. Once we had done so, we could adjoin those 51 students to the our dataset of 100 from the training and testing of our dataset. In total, we now have 151 students with sleep labels. If more time were available, we would run the smoothing algorithms for all of the students in the prediction dataset in order to generate more comprehensive observations.

The first trend we considered was the relationship between the amount of sleep a student got and the amount of steps that they logged on that same day. We found the total number of steps and total number of hours of sleep each day for all 151 people over all 31 days in October. We then created a scatterplot to visualize this distribution, as shown in Figure 4. There appears to be a weak, slightly negative relationship between the steps and sleep variables. The correlation between these variables is -0.224. This is likely due to how the sleep labels are calculated. Students who take lots of steps are marked awake during the intervals in which they are taking steps, leading more steps to be correlated with less sleep.

A common misinterpretation of sleep among college students is that a student can make up the
sleep they do not get during weekdays, on weekends. We decided to test this by creating a series of boxplots to display the amount of sleep students get on each day of the week, cited as Figure 5. We used all 151 students’ sleep data. Upon inspection, the distribution of sleep seems not to vary across the days of the week. To further test this hypothesis, we conducted a two-sample t-test to test if the mean hours of sleep a student gets on weekdays is statistically significantly different from the mean hours of sleep a student gets on weekends, at an alpha level of 0.05. Our t-test gave a p-value of 0.58, leading us to fail to reject our null hypothesis that the true mean hours of sleep a student gets differs on weekdays and weekends. With this in mind, it is important to encourage students to get a proper amount of sleep each night, as it seems unlikely that students would either be able or inclined to "make up" for poor sleep during the week with more and/or better sleep over the weekends.

5 Conclusion

5.1 Strengths and Weaknesses

The model’s strengths include its flexibility and low data requirements: the model can effectively predict sleep patterns based solely on time and steps, and can also produce probabilities rather than only classifications, indicating the extent to which the model is confident. The use of logistic regression also enables us to more carefully examine the model’s behavior and interpret its findings (e.g. by reading off coefficients). Finally, the logistic regression also gives measures of statistical significance that are useful in determining the extent to which the model’s predictions are based on a statistically significant correlation. These give marked benefits over a black-box machine learning algorithm that would afford little in the way of confidence measures.
The model certainly has a number of crucial weaknesses, however. The use of logistic regression required manipulation of the data in order to achieve the necessary independence assumption, and these manipulations substantially reduced the volume of training data available. Secondly, the assumption of a linear relationship between steps and log-odds of being asleep is, while not unreasonable, not necessarily supported by the data, particularly in the afternoon times where we observed high p-values for the correlation between log-odds and step count. Perhaps the most glaring flaw is that the model itself is composed of 96 models, each of which are not related. This is despite the fact that there is obviously a relationship between an individual’s sleep status in consecutive 15-minute intervals. Because our 96 models did not account for this relationship, smoothing was necessary to produce reasonable results.

Partly due to the construction of the model and partly due to inherent issues with predicting sleep based solely on step count, our model cannot account for sleep during the day. Furthermore, as discussed when validating the model, the model tends to overestimate sleep, likely because students tend to be inactive in the minutes before they go to bed and after they wake up; given the limited predictor variables, the model cannot differentiate between these periods of wakeful inactivity and periods of actual sleep.

Finally, the fact that the model is computationally slow to run (at least for the data preparation and smoothing portions of the model) means that it is, as currently implemented, not practical for analyzing truly large datasets unless substantial computational power is available.
5.2 Future Work

Had we more time, our first priority would be to run the smoothing algorithm for more of the prediction dataset, so that our analysis would extend more accurately to the entire given dataset rather than merely a small subset.

We would also want to spend more time examining and tailoring the smoothing algorithm, which is currently rather crude. Given how long it takes to run the algorithm, we did not have time to properly experiment to find the optimal behavior for the algorithm. While this may not be possible given the construction of our current model, we would also want to explore ways to build the dependency between sleep status in adjacent time intervals into the logistic models themselves, rather than having the dependency enforced by a smoothing algorithm that modifies the outputs of the logistic models. This would make a smoothing algorithm unnecessary, which would be preferable both in the interest of computational efficiency and because, as noted, the smoothing algorithm modifies the output of the logistic models and hence decreases the value we would otherwise get by having easy-to-interpret logistic models instead of black-box models.
6 Letter to President and Wellness Center

November 14, 2021
Re: TriangleWell Sleep and Activity Analysis

To whom it may concern:

Student health is a necessary prerequisite to a happy and productive campus. The TriangleWell initiative is a great way to encourage the spirit of physical activity and effective sleep among students. We were charged with designing a model to predict sleep habits using pedometer data of college students.

The model was designed using data from 100 students who both wore step monitoring devices and reported when they slept, either manually or through another form of technology. Our model allows us to predict whether a student is asleep or awake during a 15-minute interval based solely upon the number of steps taken during the interval and the time of day.

We applied this model to pedometer data from students who did not report sleep habits in order to predict each student’s sleep habits. This allows us to solve the issue the TriangleWell developers have been struggling with: Once we are able to predict sleep labels from a student’s steps, we can use their steps and sleep habits to learn more about the relationships between activity and sleep in college students.

We are confident in our model and its results. However, it does have some weaknesses, and we do want to note in particular that it tends to over-estimate the amount of sleep students get overall. While we encourage you to consider the model’s in developing action plans for your student wellness programming, we also consider it our duty to warn that our model and its predictions should be taken not as irrefutable fact but rather as well-supported guidance.

Our data show that, all else remaining equal, as steps increased, the average amount of sleep a student got that day slightly decreased. Keep this in mind when encouraging students to get a good night’s sleep and also remain active. According to our model, often large amounts of presumably low-quality sleep is associated with low levels of physical activity. (As a student myself, I can attest to this: large amounts of sleep often lead to a feeling of malaise during the day, and often occur on dayspreceding and following low-activity days.) Thus, it’s important to encourage students to prefer quality over quantity when it comes to sleep, and to emphasize the role that physical activity can play in getting high-quality sleep. Additionally, there does not seem to be any difference in the quantity of sleep students get throughout the week, and in particular, on weekdays versus weekends. Therefore, it’s important to promote consistent sleep, as students do not seem able to sleep any more on weekends than they can on weekdays.

Thank you for your efforts in promoting wellness among college students, and in particular for your focus on fostering healthy sleep habits. We hope our findings prove helpful in your future work on this matter, and we would be happy to answer any questions you may have, or to conduct any further analysis you may find helpful.

Sincerely,
The TriangleWell Modeling Team
7 Appendix

Our code is available on GitHub here.