Cycling performance prediction based on cadence analysis by using multiple regression

Sukhairi Sudin¹², Azizi Naim Abdul Aziz¹, Fathinul Syahir Ahmad Saad¹², Nurul Syahirah Khalid¹, Ismail Ishaq Ibrahim¹

¹Faculty of Electrical Engineering Technology, Universiti Malaysia Perlis, Kampus Pauh Putra, 02600 Arau, Perlis, Malaysia
²Centre of Excellence for Advanced Sensor Technology, Universiti Malaysia Perlis, Kompleks Pengajian Jejawi 2, 02600 Arau, Perlis, Malaysia
sukhairi@unimap.edu.my, azizinaim1998@gmail.com

Abstract. This project examined the influence of the cadence, speed, heart rate and power towards the cycling performance by using Garmin Edge 1000. Any change in cadence will affect the speed, heart rate and power of the novice cyclist and the changes pattern will be observed through mobile devices installed with Garmin Connect application. Every results will be recorded for the next task which analysis the collected data by using machine learning algorithm which is Regression analysis. Regression analysis is a statistical method for modelling the connection between one or more independent variables and a dependent (target) variable. Regression analysis is required to answer these types of prediction problems in machine learning. Regression is a supervised learning technique that aids in the discovery of variable correlations and allows for the prediction of a continuous output variable based on one or more predictor variables. A total of forty days' worth of events were captured in the dataset. Cadence act as dependent variable, (y) while speed, heart rate and power act as independent variable, (x) in prediction of the cycling performance. Simple linear regression is defined as linear regression with only one input variable (x). When there are several input variables, the linear regression is referred to as multiple linear regression. The research uses a linear regression technique to predict cycling performance based on cadence analysis. The linear regression algorithm reveals a linear relationship between a dependent (y) variable and one or more independent (y) variables, thus the name. Because linear regression reveals a linear relationship, it determines how the value of the dependent variable changes as the value of the independent variable changes. This analysis use the Mean Squared Error (MSE) expense function for Linear Regression, which is the average of squared errors between expected and real values. Value of R squared had been recorded in this project. A low R-squared value means that the independent variable is not describing any of the difference in the dependent variable-regardless of variable importance, this is letting know that the defined independent variable, although meaningful, is not responsible for much of the variance in the dependent variable's mean. By using multiple regression, the value of R-squared in this project is acceptable because over than 0.7 and as known this project based on human behaviour and usually the R-squared value hardly to have more than 0.3 if involve human factor but in this project the R-squared is acceptable.

1. Introduction

Cycling performance can be related to our performance on cadence and speed. The cadence of a cyclist can be said as the rate at which the cyclist pedals. It is the number of pedal revolutions per minute (RPMs). Increasing our cadence will improve our cycling efficiency thus allow to cycle longer and faster. Simply cadence is the speed at which we pedal. Average cyclist pedals at about 60 rpm, but
amateur and advanced cyclists pedal at much higher cadences from at least 80 rpm to more than 100 rpm. Speed, power, and also heart rate involves in the study act as features that link with the cycling performance prediction. The higher power-producing while cycling will generate more appropriate cadence. When increasing the cadence, the higher the elevation of heart rate will be thus the percentage of calories loss increase [1]. Garmin Edge 1000 is a cycling computer device with ANT+ and Bluetooth Smart which can connect to any external sensor such as a heart rate strap, power meter, speed sensor and cadence sensor. The Garmin Edge 1000 can automatically upload completed rides when in the range of a paired smartphone [2]. Cycling is a low-impact, safe activity that can be enjoyed by people of all ages, from toddlers to seniors. It's also entertaining, inexpensive, and environmentally friendly. Riding your bike to work or the store is one of the most time-efficient ways to incorporate physical exercise into your daily routine [3]. Researchers are currently trying to find a solution to the problem faced by a cyclist, especially by the novice cyclist in the field of cycling. In cycling, most novice cyclists have difficulty knowing the level of performance of the cycling during cycling activities. There are always some specific objectives behind every study and this study focused on predicting the cycling performance based on cadence analysis which linear regression algorithm been used for analysis by using Garmin Edge 1000 including with sensors provided. At first, recorded and integrate cyclist data are taking place and recorded into the database by using Garmin Connect Application. Then, from the relationship between the dependent and independent variables, the best-fit regression line could be found through regression analysis.

2. Related Work

2.1. Multiple Regression

Multiple regression is a machine learning approach that uses two or more predictors to predict a dependent variable. Multiple regression has several real-world applications in three issue domains, evaluating variable associations, producing numerical predictions, and predicting time series. To test if the relationship between the response and each term in the model is statistically significant, compare the p-value for the term to your significance threshold. The null hypothesis states that the coefficient of the term is equal to zero, indicating that there is no relationship between the term and the answer [4].

2.2. Comparison Multiple Regression with Artificial Neural Network

From [5], the author stated that the comparison findings indicated that with a lower probability threshold of 0.0001 of the mean absolute deviations from the regression line (as an indication of the estimate deviation) in the regression technique substantially outperformed the artificial neural network method. This implies that the amount of error in the estimation using the regression technique was more than the error in the artificial neural network method, implying that the neural network for estimating the barely genotypes yield was more effective than the regression approach. The correlation between the two estimates for every x in the artificial neural network models and multiple linear regression, on the other hand, was substantial and positive. But, Linear Regression is a fairly simple technique that may be readily applied to get good results. Furthermore, when compared to other sophisticated methods, these models can be trained quickly and efficiently even on systems with limited processing capacity. When compared to other machine learning methods, linear regression has a significantly reduced time complexity. Multiple regression mathematical equations are also quite simple to learn and interpret. Multiple regression also can determine the relative influence of one or more predictor variables on the criterion value. It also can find the most and least correlation between the variables. As a result, regression is relatively simple to learn and chose to apply in this study.

2.3. Residuals in Regression

Residuals play important role in regression analysis. Residuals are computed from available data and estimates as the model error to validate whether the model is good or not. If the error of the regression
model satisfies the four assumptions which are random variables with the expected value of \( o \), variance is similar to all values \( x \), the values are independent and normally distributed, then the model is count as valid\[6\]. If the assumptions regarding the error term are satisfied, so the residual plot will consist of a horizontal band of points but if not satisfied, it needs to use other ways to improve the model and obtain better results.

2.4. R-squared value

R-squared is the most important part of multiple regression. The studied R-squared metric measures how effectively a regression model predicts responses to observations. It is a statistical metric that indicates the proportion of the variation explained by an independent variable or factors in a regression model for a dependent variable. Whereas correlation describes the strength of the link between an independent and dependent variable, R-squared explains how well one variable's variation explains the variance of the other. So, if the R-squared of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs. But, R-squared cannot be used to identify whether the coefficient estimations and forecasts are skewed, thus the residual plots must be evaluated. R-squared does not show if a regression model fits your data well. The R squared value of a good model might be low. A bias model, on the other hand, can have a high R squared score. These are some of the limitations of R-squared value.

2.5. Garmin Edge 1000

Device Garmin Edge plays a major role in this study and acts as a collecting data device. The Garmin Connect Mobile app for the Edge 1000 has a variety of linked functions including live tracking, incoming call and text notifications, social media sharing, weather, wireless uploads, and sending or receiving courses and segments. When your ride is finished, the data may be instantly transmitted to Garmin Connect through Bluetooth or Wi-Fi.\[7\]. Easy to use in all weathers which would not have to worry about putting these to your bars with a Quadlock or something similar because they are waterproof. The connectivity and sensors are improved. When calibrated, the Edge 1000 uses the sensor to calculate speed and distance rather than GPS, making it more precise.

3. Methodology

3.1. Regression Line Algorithm

Regression analysis is a statistical technique for modelling the connection between a dependent (target) and independent (predictor) variable and one or more independent variables. Regression analysis, in particular, enables us to understand how the value of the dependent variable changes in relation to an independent variable while the other independent variables are maintained constant. It forecasts continuous and real data such as temperature, age, income, and price, among others\[8\].

3.2. Google Colab

Google Colab use as a platform to analyse the actual and prediction of cyclist performance. In the Colab netbook, the drive is imported and mounted at Colab. The source code will be required to proceed to the next step as it requires any Google drive account to be embedded in the notebook. After that, the import of NumPy, pandas are required to enable the data to be analysed.

The CSV file is imported to the Colab notebook and the dataset is shown in the Colab. The dataset has been separated into two groups, which are independent variables and dependent variables. Then, from the data \( X \) and \( y \), it has been separated into training and test by import the Scikit-learn model. After that, the Linear Regression been to create and the model training using the training sets. The value of the slope is get by finding the coefficient of the model. Then, the value of \( C \), intercept been print and the prediction value of \( y \) being found. The value of actual and predicted been plotted by using scatter plot graph and the regression line is applied. Finally, the value of R squared is find by getting the
model score of the model. The value of R squared is analysed whether the model explaining much in the variation of the dependent variable.

3.3. Equation

\[ y = B_0 + B_1 x + B_2 x \]  

- \( B_0 \) represents the interception
- \( B_1 \) represents the coefficient
- \( x \) represents the independent variable
- \( y \) represents the dependent variable

\( B_1 \) is the estimated regression coefficient in the multiple linear regression equation that quantifies the relationship between the risk factor \( X_1 \) and the outcome after adjusting for \( X_2 \) (\( B_2 \) is the estimated regression coefficient that quantifies the association between the potential confounder and the outcome). The basic linear regression model's \( B_1 \) is compared to the multiple linear regression model's \( b_1 \). As a rule of thumb, \( X_2 \) is considered a confounder if the regression coefficient from the simple linear regression model changes by more than 10%. Once a variable has been identified as a confounder, multiple linear regression analysis can be used to evaluate the relationship between the risk factor and the outcome while accounting for the confounder. After accounting for one or more confounding variables, the test of significance of the regression coefficient associated with the risk factor can be used to determine if the link between the risk factor is statistically significant.

4. Experimental Result

All the data collected is exported to a CSV file from the Garmin Edge. The saved CSV file in the desktop will be imported to Google Colab Notebook online and Scikit-learn to generate the analysis. A library such as “pandas” and “matplotlib.pyplot” is required to store the arrays and visualize data. The multiple regression analysis was applied in this study and the model score been observed.

![Figure 1. Statistical Table](image)

| AvgCadence | Avg HR | Avg Speed | Avg Power |
|------------|--------|-----------|-----------|
| count      | 40.000000 | 40.000000 | 40.000000 | 40.000000 |
| mean       | 71.000000 | 158.950000 | 19.060000 | 77.650000 |
| std        | 4.65373 | 9.948457 | 1.878641 | 8.033743 |
| min        | 56.000000 | 128.000000 | 13.400000 | 58.000000 |
| 25%        | 67.000000 | 153.000000 | 17.675000 | 74.000000 |
| 50%        | 72.000000 | 156.000000 | 19.250000 | 79.000000 |
| 75%        | 74.250000 | 164.000000 | 20.200000 | 82.500000 |
| max        | 77.000000 | 181.000000 | 21.500000 | 89.000000 |

Forty set of data have been used like illustrated in Figure 1 can be observed the mean of average for cadence is 71 in this project while for the mean of heart rate is 158.95, average speed mean is 19.06 and for power the mean is 77.65. The mean is useful for predicting future results when there are no extreme values in the data set. However, the impact of extreme values on the mean may be important and should be considered. Next, the standard deviation for each variables also can be observed through this statistical table. Standard deviation (std) from the statistical table above measures the spread of data distribution and it measures the typical distance between each data point and the mean. Basically, the phrase "standard deviation" refers to the amount of variability or dispersion around an average in statistics. It is a technical term for a measure of volatility. The discrepancy between the actual and average value is known as dispersion. The standard deviation increases as the dispersion or variability increases. From the standard deviation shown in the table, speed showed the lowest value of standard
deviation which explain the data are clustered closely around the mean and more reliable compared to heart rate which have the highest standard deviation which shown the data is widely spread and less reliable.

![Figure 2. Scatter Matrix Plot](image)

Figure 2 above shows the scatter matrix plot of the four datasets. It visualizes the bivariate relationships between combinations of the variable. Each scatters plot in the matrix shows the relationship between a pair of variables to allow many relationships to be explored in one chart. From the plot above, we can see the relationship of average cadence (AvgCadence) with average heart rate (Avg HR) have a negative correlation as it goes from high-value of y-axis down to high-value of x-axis which will give the value of the slope is negative which is -0.12756. The heart rate involvement in the performance can prove whether the performance is good or bad. If the heart rate at high while the cadence at low, the performance can be said as not good. If the heart rate at low while the cadence at high, the performance of the cyclist can be said as good. From the plot, the heart rate has a negative correlation because heart rate is physiological data. There are many factors that can affect the data compared to power and speed. Physiological data involve direct or indirect observation related to the system and subsystem of the human body. For example, the cyclist may have a high heart rate due to his emotions or feeling at that time. For the relationship between average cadence with average power (Avg Power) and average speed (Avg Speed), both have a positive relationship as it is from the origin out to high x-value to y-value which the value of slope for average power and average speed are 0.01366 and 1.94119. Both are directly proportional to the cadence.

![Figure 3. Regression Line](image)

From the regression line illustrated in Figure 3, it can be observed that there is some value of prediction that deviated from the actual value. This can be stated as a prediction error for the analysis. From the graph shown above, the prediction was mostly underestimating the actual outcome as (y <
x). The deviation difference between the actual and predicted value is ranges between 0.5 to 2.5. The value of y interception which indicates as constant in the analysis is 52.92769155. This value indicates a constant value. The intercept is the expected mean value of y when all x = 0. This show the model is intrinsically linear regression.

![Figure 4. Residual Plot](image)

The red colour of the graph line in Figure 4 indicates the actual value or testing value while the blue colour of the graph indicates the predicted values. The curve depicts a density plot, which is effectively a smoothed-out version of a histogram. The y-axis is in terms of density, and the histogram is automatically normalised to have the same y-scale as the density plot. From this plot, it tells the residuals are quite homoscedastic as the residuals. Homoscedasticity refers to a condition in which the variance of the residual term is constant or nearly so [9].

From the model, the value of R-squared is 0.85. The 85% of R-squared indicates the cycling performance moves relatively in line with the index. From the value, it shows that the model explains 85% of the variability of the response data around the mean which means quite good.

5. Conclusion

After interpretation of data, it can conclude that the value of R squared from the regression line quite satisfy as 0.85 is nearly to 1 and can be said acceptable. From author [10], can be remarks by having high value of R squared are more acceptable to act as predictor. But, the model is not perfect due to many factors. From the matrix plot, the power and speed have positive correlation with the cadence while heart rate show negative correlation. It can conclude that heart rate may affect by others factor such as weather, humidity, emotions, dietary, and other factors that directly and indirectly related to the human body as heart rate known as physiological data. For cadence, power and speed, they are known as kinetic and kinematic data which involves displacement and orientation of body segments only. It depends on the power and strength of the cyclist and therefore they have positive correlation. From the results, it can be observed that the prediction was mostly underestimating the actual outcome as (y < x). It also showed that the model need to be improve as the R-squared value not completely explaining much the dependent variable. The difference between actual and predicted value also explained that there are error between them.

References

[1] “High Cadence Cycling and Speed - Will High Cadence Make You Faster?” https://www.bicycling.com/news/a26305270/high-cadence-cycling-not-increase-speed/ (accessed Jul. 05, 2021).

[2] A. F. Maio and J. A. Afonso, “Wireless cycling posture monitoring based on smartphones and bluetooth low energy,” in Lecture Notes in Engineering and Computer Science, 2015, vol. 2217, pp. 653–657.

[3] “Cycling - health benefits - Better Health Channel.” https://www.betterhealth.vic.gov.au/health/healthyliving/cycling-health-benefits (accessed Jul.
[4] C. Flatt and R. L. Jacobs, “Principle Assumptions of Regression Analysis: Testing, Techniques, and Statistical Reporting of Imperfect Data Sets,” Adv. Dev. Hum. Resour., vol. 21, no. 4, 2019, doi: 10.1177/1523422319869915.

[5] S. Nakagawa and H. Schielzeth, “A general and simple method for obtaining R2 from generalized linear mixed-effects models,” Methods Ecol. Evol., vol. 4, no. 2, 2013, doi: 10.1111/j.2041-210x.2012.00261.x.

[6] “Statistics - Residual analysis | Britannica.” https://www.britannica.com/science/statistics/Residual-analysis (accessed Jul. 06, 2021).

[7] “Review: Garmin Edge 1000 performance bundle | road.cc.” https://road.cc/content/review/137336-garmin-edge-1000-performance-bundle (accessed Jul. 05, 2021).

[8] C. Flatt and R. L. Jacobs, “Principle Assumptions of Regression Analysis: Testing, Techniques, and Statistical Reporting of Imperfect Data Sets,” Adv. Dev. Hum. Resour., vol. 21, no. 4, 2019, doi: 10.1177/1523422319869915.

[9] “Homoscedasticity - Wikipedia.” https://en.wikipedia.org/wiki/Homoscedasticity (accessed Jul. 06, 2021).

[10] Sudin, Sukhairi & md shakaff, ali yeon & Zakaria, Ammar & Salleh, Ahmad & Kamarudin, Latifah & Azmi, Noraini & Ahmad Saad, Fathinul. (2018). Real time track cycling performance prediction using ANFIS system. International Journal of Performance Analysis in Sport. 18. 10.1080/24748668.2018.1523520.