Impact of Variable Transformations on Multiple Regression Models for Enhancing Gait Normalization

FLORA, FERREIRA*
Centre of Mathematics, School of Sciences, University of Minho
fjferreira@math.uminho.pt

JHONATHAN, Barrios
Centre of Mathematics, School of Sciences, University of Minho
jhonathanbarrios21@gmail.com

PAULO, Barbosa
Centre of Mathematics, School of Sciences, University of Minho
pjbarbosa98@gmail.com

MIGUEL, F, Gago
Neurology Department, Hospital da Senhora da Oliveira; and School of Medicine, Life and Health Sciences Research Institute (ICVS), University of Minho
miguelgago@hospitaldeguimaraes.min-saude.pt

ESTELA, Bicho
Algoritmi Research Centre, School of Engineering, University of Minho
estela.bicho@dei.uminho.pt

WOLFRAM, Erlhagen
Centre of Mathematics, School of Sciences, University of Minho
wolfram.erlhagen@math.uminho.pt

ABSTRACT

Gait analysis has become an important tool in clinical practice for monitoring disease progression and evaluating therapeutic interventions. However, a subject's gait characteristics can be affected by physical characteristics such as age and height, which can interfere with accurate comparisons between subjects. MLR normalization has been shown to be effective in reducing interference from subject-specific physical properties, but non-linear effects can still impact the results. In this study, the independent variables were transformed to improve normalization performance, and the results indicate that using MR normalization with data transformation can effectively de-correlate physical characteristics from gait variables, improving the model fit and augment the capability to compare subjects with varying physical characteristics. This study provides valuable insights into the use of MLR models for gait normalization, with potential applications in clinical practice and research.

CCS CONCEPTS

• **Linear Regression models**; • **Data normalization**; • **Multivariate statistics**;

KEYWORDS

Gait analysis, Multiple linear regression models, Data transformation, Gait normalization

ACM Reference Format:
FLORA, FERREIRA*, JHONATHAN, Barrios, PAULO, Barbosa, MIGUEL, F, Gago, ESTELA, Bicho, and WOLFRAM, Erlhagen. 2023. Impact of Variable Transformations on Multiple Regression Models for Enhancing Gait Normalization. In *2023 6th International Conference on Mathematics and Statistics (ICoMS 2023), July 14–16, 2023, Leipzig, Germany*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3613347.3613363

ICoMS 2023, July 14–16, 2023, Leipzig, Germany

ACM ISBN 979-8-4007-0018-7/23/07.
https://doi.org/10.1145/3613347.3613363

1 INTRODUCTION

Gait analysis has become an increasingly important tool in clinical practice due to its potential to provide an objective assessment of gait impairment [1]. Gait analysis involves measuring various spatiotemporal parameters related to an individual's walking pattern, such as stride length, step width, cadence, and walking speed. These parameters can be used to monitor disease progression, evaluate therapeutic interventions, and identify gait abnormalities in various conditions, such as Parkinson's disease [2].

However, a subject's gait characteristics can be affected by various factors, such as age, height, weight, and walking speed, which can interfere with accurate comparisons between individuals [3, 4] . To address this issue, researchers have employed various normalization techniques to reduce the impact of subject-specific physical properties on gait measures [5, 6]. Among these techniques, multiple linear regression (MLR) normalization has been shown to be effective in reducing interference from subject-specific physical properties and their gait variables, in comparison to other methods like dimensionless equations and detrending techniques [7]. However, MLR models are limited in their ability to capture non-linear effects [8], which can impact the accuracy of gait normalization.

To address this limitation, independent variables can be transformed to improve normalization performance, reduce correlations between subject-specific physical characteristics and gait features, and decrease the dispersion of gait data between subjects during walking. Transformations such as logarithm, square root, square, or cube can improve the model fit and correct violations of the model assumptions [9]. In this study, the independent variables were transformed using various mathematical functions to assess the impact of independent variable transformations on MLR models for gait normalization.

Previous studies have identified age, height, weight, sex, walking speed, and stride length as variables significantly affecting gait data

[3, 10–14]. Therefore, as in [6], these variables were considered independent variables in this study. The proposed approach involves transforming the independent variables using various mathematical functions, such as natural logarithm, square root, square, and cube, and comparing the performance of MLR models before and after the transformations.

The paper is structured as follows: Section 2 describes the gait data and the proposed gait normalization approach. Section 3 presents the results and discussion, including the impact of the independent variable transformations on MLR models for gait normalization. Section 4 outlines the conclusions and future directions for this research, emphasizing the potential of using transformed independent variables in MLR models for improving gait normalization accuracy in clinical practice and research.

2 MATERIALS AND METHODS

2.1 Gait Data

This study utilized gait data from 36 healthy adults with ages ranging from 20 to 85 years, weights ranging from 48.9 to 94.2 kg, and heights ranging from 1.53 to 1.89 meters, as reported in a previous study [6]. To collect gait data, two Physilog®sensors (Gait Up®, Switzerland) were attached to the dorsum of each shoe using elastic bands. The participants were instructed to walk a 60-meter continuous course, which consisted of a 30-meter corridor with one turn, at a self-selected speed.

Various spatial, temporal, and foot clearance gait variables were assessed from the participants' gait data, including: speed (velocity of one stride), cycle duration (duration of one stride), cadence (number of strides per minute), stride length (distance between successive initial ground contacts using the same foot), stance (percentage of stride that the foot is on the ground), swing (percentage of stride that the foot is in the air), loading (percentage of stance between the heel strike and the foot placed fully on the ground), foot flat (percentage of stance where the foot is fully on the ground), pushing (percentage of stance between the foot fully positioned on the ground and the toe leaving the ground), double support (percentage of stride that both feet touch the ground), peak swing (maximum angular velocity during swing), strike angle (angle between the foot and the ground when the heel hits the ground), lift-off angle (angle between the foot and the ground when the toes are leaving the ground), maximum heel (maximum height above the ground reached by the heel), maximum toe clearance 1 (maximum height above the ground reached by the toes after maximum heel strike), minimum toe clearance (minimum height of the toes during the swing phase), and maximum toe clearance 2 (maximum height above the ground reached by the toes just before heel strike).

2.2 Multiple Regression Normalization with data transformations

The MR Normalization with data transformation was carried out through the following steps:

- Scatter plots were created to visually examine the relationships between variables.
- Independent variables (X) underwent quadratic (X^2), cubic (X^3), square root ($\sqrt{X}$), inverse ($1/X$), natural logarithm ($ln(X)$, and natural base elevation to a power (e^X) transformations. This resulted in new predictors that were functions of the existing variables. Pearson correlation coefficients were used to assess the associations between the original and new predictors with the gait variable. Predictors with a significant Pearson coefficient correlation at the 0.10 level were selected for the next step.
- Regression models were developed for all possible combinations of the selected predictors. Variance inflation factors (VIFs) were calculated to ensure that multicollinearity was absent from the models. Any models with a VIF greater than 3.3 were excluded [15].
- Akaike's information criterion (AIC) and adjusted R^2 metrics were used to identify the best-fit model. Only the models that had significant variables were taken into consideration. For each best-fit regression model, a normal quantile-quantile (Q-Q) plot of the residuals and a residual plot were generated to verify the assumptions of normality of regression residuals and homoscedasticity, respectively. Standardized residual values were assessed to identify influential outliers.
- The best-fit regression models were used to normalize each gait variable by dividing the value of the original dependent gait variable y_i, by the predicted gait variable $\widehat{y_i}$. This results in a new value, y_i^N, which represents the normalized gait variable for the ith observation within each subject.

The linear regression model's effectiveness in de-correlating gait parameters through normalization was evaluated using both Pearson (linear) and Spearman (ranking-based) correlation coefficients before and after normalization.

Additionally, the subject group was randomly split into two subgroups with age, height and speed differences, and the Mann-Whitney U Test was employed to examine differences between them. This was done to evaluate the capability of MR normalization with and without data transformations to facilitate the comparison of gait variables between cohorts with diverse physical characteristics.

3 RESULTS AND DISCUSSION

3.1 Multiple linear regression models

Table 1 displays the results of multiple linear regression models with and without data transformation, including the independent variables, AIC, and adjusted R^2 values. Comparison of the AIC and adjusted R^2 values indicates that data transformation improves the model fit, particularly for cycle duration, cadence, foot flat, pushing, double support, peak swing, strike angle, and maximum heel clearance. Foot flat was the only gait variable that did not show improvement in these two metrics. Interestingly, no significant Pearson coefficient correlation at the 0.10 level was found between all speed variables (original and transformed) and foot flat, and therefore, the speed (or other variables obtained by transformation) were not included in the final model. The lack of significant correlation between speed variables and the dependent variable suggests that speed may not be a crucial factor in predicting foot flat. For maximum Toe Clearance 1, no significant Pearson coefficient correlation at the 0.10 level was found for all original and transformed independent variables. As per the proposed approach, gait MR normalization is not necessary in this case.

Table 1: Independent variables present in the multiple regression models without data transformation [6] and with data transformation, for the gait variables. The adjusted (Adj) *R* squared and Akaike information criterion (AIC) are shown.

Gait variable	MR normalization			MR normalization with data transformation		
	Independent variables	AIC	Adj R square	Independent variables	AIC	Adj R square
Temporal Variables						
Cycle duration	$A,\ H,\ V$	-143.7	0.771	$A^3,\ H^3, ln\ (V)$	-144.8	0.803
Cadence	$A,\ H,\ V$	197.8	0.776	$A^3,\ H^3,\ \sqrt{V}$	193.3	0.802
Stance	V	146.6	0.227	$\frac{1}{V}$	148.1	0.238
Swing	V	146.6	0.227	$\frac{1}{V}$	148.1	0.238
Loading	$V,\ S,\ W$	154.3	0.363	$S, \sqrt{W}$	154.4	0.346
Foot Flat	$A,\ V,\ W$	190.0	0.662	$\frac{1}{A}\ ,\ W^2, \frac{1}{V}$	183.7	0.717
Pushing	$A,\ V,\ W$	181.5	0.600	$\frac{1}{A}\ , ln\ (V)$	178.5	0.624
Double Support	V	187.9	0.227	$\frac{1}{V}$	186.3	0.308
Spatial Variables						
Stride Length	$A,\ H,\ V$	-126.3	0.928	$A^3,\ H^3,\ \sqrt{V}$	-130.9	0.937
Peak Swing	V	340.0	0.570	$\frac{1}{V}$	337.7	0.598
Foot Clearance Variables						
Strike Angle	$H,\ S,\ SL$	195.2	0.515	$\frac{1}{A}, \sqrt{H},\ S, \frac{1}{SL}$	186.3	0.560
Lift-off-Angle	$A,\ SL$	214.5	0.767	$A^3, ln\ (SL)$	211.9	0.783
MaxHC	$A,\ H,\ S$	-147.3	0.449	$A^3,\ S$	-151.8	0.501
MaxTC1	SL	-170.2	0.006	-	-	-
MinTC	A	-244.6	0.426	$\sqrt{A}$	-246.8	0.427
MaxTC2	$A,\ S,\ SL$	-180.9	0.734	$\sqrt{A},\ S,\ SL$	-181.0	0.735

MaxHC: Maximum Heel Clearance; MaxTC1: Maximum Toe Clearance 1; MinTC: Minimum Toe Clearance; MaxTC2: Maximum Toe Clearance 2. The independent variables are age (A), height (H), speed/velocity (V), sex (S), weight (W) and stride length (SL).

Although the sample size of 36 subjects in this study may limit the accuracy of the regression models obtained in [6] the results show that the multiple linear regression models developed are comparable in their ability to accurately predict outcomes to previously published regression models, such as those presented in [3, 12] . This suggests that the improvement observed by applying data transformation may be observed in other samples as well.

3.2 Decorrelation through normalization

Table 2 presents the Spearman coefficients correlations between physical characteristics, speed, and stride length with gait variables. Normalization with MLR without data transformation of height and weight resulted in a significant Spearman coefficient correlation with peak swing and strike angle, respectively. However, the MLR with transformation proved to be better not only in improving the model fit but also in de-correlating physical characteristics from gait variables.

The MR normalization approach effectively de-correlated data from physical characteristics, speed, and stride length, reducing the correlation coefficient from $|\rho| < 0.87$ to $|\rho| < 0.46$. Additionally, MR normalization with data transformation further reduced the correlation coefficient to $|\rho| < 0.31$. This suggests that using data transformation during MR normalization is more effective in eliminating the influence of physical characteristics on gait variables, resulting in a better model fit.

3.3 Gait Differences between Two groups of Healthy Subjects

Table 3 summarizes the subject demographics and walking speeds of the two groups randomly split. No statistically significant differences were found for weight and sex. It should be noted, however, that the group was divided randomly while ensuring that there was a statistically significant difference in age, weight, and speed between the groups.

Table 4 provides a detailed comparison of gait variable values between Group 1 and Group 2 in their raw form, as well as after being normalized using MR normalization alone and MR normalization with data transformation approaches. The statistical analysis revealed significant differences in cycle duration, cadence, pushing, stride length, lift-off angle, maximum heel, minimum toe clearance, and maximum toe 2 between the two groups. These differences can primarily be attributed to variations in age, height, and walking speed, as both groups were composed of healthy individuals. However, after applying normalization using either approach, no

Table 2: Spearman correlation coefficients (ρ) for the data before (raw), after normalization using multiple linear regression models without data transformations (NL), and after normalization using multiple linear regression models with data transformations (NLT)

Gait Variable	Age			Height			Weight			Sex			Speed			Stride length		
	RAW	NL	NLT	RAW	NL	NLT	RAW	NL	NLT	RAW	NL	NLT	RAW	NL	NLT	RAW	NL	NLT
Cycle Duration	**-.39**	-.01	.03	**.49**	.03	-.01	**.37**	.10	.05	.24	-.10	-.01	**-.36**	.04	.06	.08	.31	.28
Cadence	**.39**	.01	-.04	**-.50**	-.04	.01	**-.37**	-.14	-.07	-.24	.11	.05	**.35**	-.02	-.03	-.08	-.29	-.25
Stance	.16	.02	.02	-.05	.08	.07	.13	.16	.13	-.06	-.01	-.02	**-.34**	.08	.09	-.31	.07	.07
Swing	-.16	-.05	-.03	.05	-.07	-.06	-.13	-.14	-.13	.06	.01	.03	**.34**	-.09	-.08	.31	-.07	-.05
Loading	-.19	-.03	.10	.04	.13	-.31	**-.33**	.10	-.12	.30	.03	-.25	.26	-.02	-.06	.26	.07	.05
Foot Flat	**.49**	.07	.01	-.23	-.09	-.08	**.40**	.00	.01	-.16	-.21	-.24	**-.72**	-.24	-.17	**-.67**	-.22	-.19
Pushing	**-.52**	-.10	-.01	.24	.05	-.21	-.30	-.03	-.27	.09	.00	-.11	**.74**	.23	.09	**.69**	.19	.03
Double Support	.15	.01	-.01	-.03	.07	.07	.18	.08	.09	-.13	-.16	-.16	**-.41**	.07	.08	**-.35**	.06	.06
Stride Length	**-.58**	-.12	-.02	**.61**	.10	.00	.12	.10	.05	.18	-.06	-.07	**.87**	.13	.05	–	–	–
Peak Swing	-.24	.08	.00	.04	**-.33**	-.25	-.20	-.29	-.18	.09	.03	.14	**.70**	.00	.00	**.56**	-.04	-.07
Strike Angle	**-.48**	-.26	.02	.30	-.03	-.04	-.27	**-.46**	-.25	.37	-.06	-.12	**.49**	-.04	-.01	**.61**	.05	.06
Lift-Off Angle	**.75**	.05	.05	**-.51**	-.17	-.14	.05	-.11	-.15	-.26	.12	.23	**-.73**	.15	.12	**-.85**	.00	-.04
MaxHC	**-.56**	-.08	-.03	**.56**	.07	.19	.25	.01	.08	**.54**	-.02	-.04	.22	.13	.13	**.44**	.15	.16
MaxTC1	.17	.00	–	-.07	.09	–	.10	.12	–	-.01	.06	–	-.23	-.02	–	-.16	.09	–
MinTC	**.61**	-.08	-.13	**-.36**	-.06	-.02	.08	-.05	-.05	-.27	-.18	-.16	**-.39**	-.10	-.08	**-.46**	.00	.03
MaxTC2	**-.72**	.03	-.02	**.56**	-.04	-.02	-.05	-.20	-.10	.45	-.02	-.02	**.53**	-.09	-.12	**.73**	-.02	-.02

Bold values indicate Spearman correlation coefficients significant at the 0.05 level (2- tailed).

Table 3: Comparison of Physical Characteristics and walking Speed Measures between Two Randomly Split Groups

	Group 1 Median [Minimum, Maximum]	Group 2 Median [Minimum, Maximum]	*p-value*
Age (years)	70.5 [26,35]	34.5 [20, 77]	0.0008*
Height (m)	1.63 [1.53, 1.78]	1.71 [1.53, 1.89]	0.0128*
Weight (kg)	66.9 [48.9, 94.2]	70.25 [53.0, 86.5]	0.6923*
Speed (m/s)	1.24 [0.88, 1.66]	1.38 [1.17, 1.79]	0.0142*
Male (%)	50%	22.3%	0.0698+

* Mann-Whitney U test. + Fisher Exact T Test

significant differences were found except for stride length. This finding corroborates previous studies, such as [2, 3], which have demonstrated that normalizing gait variables through linear regression models can enhance the ability to compare individuals with different physical characteristics and walking speeds. Furthermore, while there was a statistical difference in stride length after MR normalization, no such difference was observed after MR normalization with data transformation, which can be attributed to the high correlation between stride length and age, height, and walking speed.

4 CONCLUSION

The study's findings demonstrate that using data transformations during MR normalization can effectively improve the de-correlation of spatial, temporal, and foot clearance gait variables from physical properties, walking speed, and stride length when compared to MR normalization alone. This improvement enhances the capability to compare individuals with different physical characteristics and walking speeds, thereby improving the accuracy of conclusions drawn about their gait patterns. It is important to acknowledge that this study is limited by the size of the dataset, potentially hindering its ability to comprehensively capture the variability in gait and physical characteristics within the population. In order to validate and generalize the findings of this study, it would be highly valuable to incorporate a larger and more diverse dataset in future research.

Previous studies, such as [2, 7], have highlighted the potential of using machine learning techniques for gait classification based on MR normalized gait variables. This approach has shown higher accuracy compared to using raw data. As a potential area for future research, exploring the impact of different data transformations on MR normalization and assessing their potential for improving the

Table 4: Comparison between the values of gait variables in Group 1 and Group 2. Mann-Whitney U test p-values are shown for raw gait data, the MR normalized gait (NL) and MR normalized gait with data transformation.

Gait Variable	*p-value*			Gait Variable	*p-value*		
	RAW	**NL**	**NLT**		**RAW**	**NL**	**NLT**
Cycle Duration	**0.0279**	0.0738	0.1101	Stride Length	**0.0003**	**0.0218**	0.0967
Cadence	**0.0279**	0.0517	0.1032	Peak Swing	0.0905	0.6693	1.0
Stance	0.2892	0.9621	0.9874	Strike Angle	0.0689	1.0	0.7397
Swing	0.2892	0.9118	1.0	Lift-Off Angle	**0.0001**	0.3038	0.3843
Loading	0.716	1.0	0.1892	MaxHC	**0.0099**	0.1329	0.1592
Foot Flat	0.0689	0.5583	0.4572	MaxTC1	0.6239	0.8868	0.8371
Pushing	**0.0498**	0.937	0.8868	MinTC	**0.0218**	0.7397	0.8371
Double Support	0.5166	0.937	0.8868	MaxTC2	**0.0005**	0.5373	0.4383

performance of machine learning-based gait classification models could be a valuable endeavor.

Overall, the findings of this study provide valuable insights into the potential use of MLR models for gait normalization in both clinical practice and research settings. By employing MR normalization with data transformation, researchers and clinicians can obtain more accurate and reliable information about gait patterns, which can aid in the diagnosis and treatment of various gait-related conditions. Furthermore, the findings highlight the potential for MLR models to be used in combination with machine learning techniques to further enhance the accuracy of gait classification and monitoring.

ACKNOWLEDGMENTS

Supported by Portuguese funds through the Centre of Mathematics and the Portuguese Foundation for Science and Technology (FCT), within the projects UIDB/00013/2020 and UIDP/00013/2020.

REFERENCES

[1] Y. Yan, O. M. Omisore, Y.-C. Xue, H.-H. Li, Q.-H. Liu, Z.-D. Nie, J. Fan and L. Wang, "Classification of neurodegenerative diseases via topological motion analysis—A comparison study for multiple gait fluctuations," IEEE Access, vol. 8, pp. 96363–96377, 2020.

[2] C. Fernandes, F. Ferreira, M. Gago, O. Azevedo, N. Sousa, W. Erlhagen and E. Bicho, "Gait classification of patients with Fabry's disease based on normalized gait features obtained using multiple regression models," in IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2019.

[3] V. Mikos, S.-C. Yen, A. Tay, C.-H. Heng, C. L. H. Chung, S. H. X. Liew, D. M. L. Tan and W. L. Au, "Regression analysis of gait parameters and mobility measures in a healthy cohort for subject-specific normative values," PloS one, vol. 13, no. 6, p. e0199215, 2018.

[4] F. Ferreira, M. F. Gago, E. Bicho, C. Carvalho, N. Mollaei, L. Rodrigues, N. Sousa, P. P. Rodrigues, C. Ferreira and J. Gama, "Gait stride-to-stride variability and foot clearance pattern analysis in Idiopathic Parkinson's Disease and Vascular Parkinsonism," Journal of biomechanics, vol. 92, pp. 98–104, 2019.

[5] J. Braga, F. Ferreira, C. Fernandes, M. F. Gago, O. Azevedo, N. Sousa, W. Erlhagen and E. Bicho, "Gait characteristics and their discriminative ability in patients with fabry disease with and without white-matter lesions," in Computational Science and Its Applications–ICCSA 2020: 20th International Conference, Proceedings, Part III, Cagliari, Italy, 2020.

[6] C. Fernandes, F. Ferreira, R. L. Lopes, E. Bicho, W. Erlhagen, N. Sousa and M. F. Gago, "Discrimination of idiopathic Parkinson's disease and vascular parkinsonism based on gait time series and the levodopa effect," Journal of Biomechanics, vol. 125, p. 110214, 2021.

[7] F. Wahid, R. K. Begg, C. J. Hass, S. Halgamuge and D. C. Ackland, "Classification of Parkinson's disease gait using spatial-temporal gait features," IEEE journal of biomedical and health informatics, vol. 19, no. 6, pp. 1794–1802, 2015.

[8] M. L. Callisaya, L. Blizzard, M. D. Schmidt, J. L. McGinley and V. K. Srikanth, "Sex modifies the relationship between age and gait: a population-based study of older adults," The Journals of Gerontology Series A: Biological Sciences and Medical Sciences, vol. 63, no. 2, pp. 165–170, 2008.

[9] D. C. Montgomery, Design and analysis of experiments, John wiley & sons, 2017.

[10] R. Senden, K. Meijer, I. Heyligers, H. Savelberg and B. Grimm, "Importance of correcting for individual differences in the clinical diagnosis of gait disorders," Physiotherapy, vol. 98, no. 4, pp. 320–324, 2012.

[11] F. Dadashi, B. Mariani, S. Rochat, C. J. Bula, B. Santos-Eggimann and K. Aminian, "Gait and foot clearance parameters obtained using shoe-worn inertial sensors in a large-population sample of older adults," Sensors, vol. 14, no. 1, pp. 443–457, 2013.

[12] F. Wahid, R. Begg, N. Lythgo, C. J. Hass, S. Halgamuge and D. C. Ackland, "A multiple regression approach to normalization of spatiotemporal gait features," Journal of applied biomechanics, vol. 32, no. 2, pp. 128–139, 2016.

[13] C. Kirtley, M. W. Whittle and R. Jefferson, "Influence of walking speed on gait parameters," Journal of biomedical engineering, vol. 7, no. 4, pp. 282–288, 1985.

[14] L. Alcock, B. Galna, R. Perkins, S. Lord and L. Rochester, "Step length determines minimum toe clearance in older adults and people with Parkinson's disease," Journal of biomechanics, vol. 71, pp. 30–36, 2018.

[15] N. Kock and G. Lynn, "Lateral collinearity and misleading results in variance-based SEM: An illustration and recommendations," Journal of the Association for information Systems, vol. 13, no. 7, 2012.