Anthropometric Data Analytics: a Portuguese Case Study

António Barata[1], Lucília Carvalho[2], Francisco M Couto[1]

[1]LaSIGE, Faculdade de Ciências da Universidade de Lisboa, Portugal
[2]Hospital de Egas Moniz, Centro Hospitalar de Lisboa Ocidental, Portugal
apbarata@gmail.com, fcouto@di.fc.ul.pt,
lcmcarvalho@chlo.min-saude.pt

Abstract. Large amounts of information are systematically generated throughout the course of scientific research and progress. In our case, observations representing the Portuguese population within the central-southern region of Portugal were collected throughout various foetal autopsy procedures. Gestational age (GA) and measured distances and weights of numerous anthropometric features and organs, respectively, were recorded per singleton (24 variables in total). This work seeks to elaborate on the accuracy of different foetal parameters in terms of GA estimation, making use of principal component analysis (PCA) and regression techniques. We created a dataset of 450 foetuses, ranging from 13 to 42 weeks of age, to compute both PCA and regression models. Initial exploratory analysis shed light onto which variables are most explanatory in terms of foetal development, and are thus most likely suitable for predictive rolls. We produced clusters of models, based on coefficient of determination (R^2) values, by comparing the squared sum of residuals between models (significance level $\alpha = 0.05$). Models comprised of linear combinations of different variables exhibited significantly higher values of R^2 (p-value ≤ 0.05) when compared to single variable models. Across all regressions, crown-heel length (CHL), crown-rump length (CRL), and foot length (FL) are constantly present within the cluster of best predictors of gestational age. Depending on the type of regression analysis applied, body weight (Body), hand length (HL) also fall onto the same category.

Keywords: Foetopathology; Foetus; Prediction; Estimate; Gestational age; Crown-rump length; Crown-heel length; Foot length;

1 Introduction

Performing rigorous estimations of gestational age is invaluable for correct diagnosis and optimum treatment of disease during the neonatal period. GA prediction is an essential tool for parental counselling and to plan for appropriate perinatal care. It is also a prime requisite for foetal autopsy, particularly in situations of criminal abortion, alleged infanticide, and medically-terminated pregnancies. Previous peer-reviewed studies have elaborated on the accuracy of different foetal parameters in gestational age prediction [1], particularly head circumference (HC), HL, FL, CRL, and CHL [2, 5]. Model analysis and hypothesis tests may help determine not only how different measurements and weights are linked to foetal developmental age, but also which variables might be classified and ordered in terms of their predictive capabilities.

In regards to anthropometric data analytics, other published papers often approach the validity of different measured variables for conceptual age estimation [6, 10], and the quantitative standards of those measurements for foetal and neo-natal autopsy [11]. Regression analysis and model fitting are widely accepted and used in this field of work, hence being viewed as reliable tools for knowledge production [12]. Other relevant publications may also be found, discussing the relationship between different methods of analysis and discriminating regression properties, enabling model validation for subsequent selection [13, 14]. Currently, the application of analytical and statistical methods for the evaluation of information is accomplished with the use of data manipulative software [15, 16]. For these computer programs to be beneficial, however, all data must be made digitally available. Without a proper data frame, analysis of data becomes tedious and/or unfeasible.

Based on foetal autopsy records, we created a dataset of 450 individuals, each comprised of 24 foetal parameters. PCA produced results indicating CHL, CRL, and FL variables as the most explanatory in terms of total data variance. By comparing regressions models, Body and HL parameters were also found to be significantly viable measurements for GA estimation. Background information regarding related work is discussed in Section 2. The following section describes the methodological approaches used, while Section 4 presents the results of said methods. Discussion of obtained results and final remarks pertain to the 5th and final Section of this paper.

2 Case Study

For several years, the foetopathology department of Hospital de Egas Moniz, has been conducting the analysis and evaluation of foetal mortality cases pertaining to the central-southern region of Portugal. Each foetal autopsy produces a physical report file containing, amongst other relevant medical information, measurements and weights of the foetus. Whenever a foetopathology instance is concluded, the file is then archived within a dossier. This type of information processing and storage does not permit direct access to harboured values in more than a few cases at a time. Reports are regarded independently of each other, making any data study laborious and time-consuming.

To address this challenge, we developed a database representing foetal autopsy records. Each report had to be manually inserted, due to discrepancies of cursive between files, excluding the use of optical character recognition (OCR) software. A total of 450 individuals between the ages of 13 and 42 inclusive were inserted into the database.

3 Methods

Given the format of each autopsy report file in this work, a database was constructed and algorithms to store, retrieve, and manipulate information were devised. Python was applied as the programming language for these tasks mainly due to its extensive libraries and packages, notably the SQLite3, NumPy, and SciPy modules [17, 19]. IBM’s SPSS software [20] was also utilized due to its inbuilt statistical applications, concretely PCA and variable selection algorithms for multiple linear regression.

3.1 Data Structure

24 quantitative variables were selected to represent each foetal autopsy case. Retrieved according to autopsy protocol, the extensive list of recorded foetal parameters follows: GA, CHL, CRL, HC, chest circumference (CC), abdominal circumference (AC), FL, HL, middle finger length (MFL), intercommissural distance (ID), philtrum length (PL), inner canthal distance (ICD), outer canthal distance (OCD), left palpebral fissure width (LPFW), right palpebral fissure width (RPFW), left ear length (LEL), right ear length (REL), body, kidneys, thymus, spleen, liver, lungs, and adrenals. Paired organs are represented by their combined weight. Units comprise of week (GA), centimetre (distances and lengths), and gram (organ and body weights). Additionally, GA values consist of observed occurrences, reported throughout every case file, and not mere value estimations.

3.2 Initial Exploration and Modelling

SPSS was used to conduct the initial PCA, which would provide foresight onto possible outcomes of successive regression models. Computed extraction communalities, loadings, explained variance per component, and adequacy parameters were consequently inspected. Computation of multiple linear regression models was performed through the same IBM software. GA was selected as the dependent variable, while the remaining 23 features were used as predictors. All available regression algorithms for variable selection (Enter, Stepwise, Remove, Backward, and Forward) were utilized and their outputs taken into consideration. Models were selected based on statistically significant coefficient values ($\alpha = 0.05$), as well as Durbin-Watson and R^2 values. Standardized and un-standardized β-weights were also a point of interest for later model comparison. In total, 5 different k^{th} degree polynomial regression functions were fit onto each of the 23 variables, for $k \in \{1, 2, 3, 4, 5\}$. Each variable dataset consisted of pairs of variable-age points, where each pair represents the gestational age and recorded variable value of a singleton foetus. The NumPy module polyfit() function was used to output each single variable model. R^2 and estimated parameter values were recorded for all regressions presenting a significant *p*-value for the null hypothesis that the estimated coefficients are equal to zero.

3.3 Model Comparison

Regression models were compared based on each model's proportion of variance in the dependent variable predictable by the independent variable. The *F*-statistic was selected and computed using the squared sum of residuals (SSR) and degrees of freedom of the models being compared [21]. A significance level of $\alpha = 0.05$ was established. The SciPy module stats.f.cdf() function was used to compute *p*-values. Each multiple linear regression model was compared to all other multiple and polynomial models. Polynomial models were compared to other polynomial models if and only if both models pertained to the same polynomial degree. The resulting *p*-values were stored for later interpretation.

4 Results

4.1 Principal Component Analysis

For our dataset, the Kaiser-Meyer-Olkin (KMO) index for sampling adequacy had a value of 0.973 while the *p*-value corresponding to the χ^2-statistic associated with Bartlett's test of homoscedasticity was below 5x10^{-4}. PCA produced only one significant component (eigenvalue $\geq$ 1) explaining 93.486% of total data variance. Communality and loading values for all variables are shown below.

Table 1. Communality and loading values per variable. Darker shades representing lower values. Table spliced due to size constraints.

	Communality	**Loading**
CRL	0.963	0.981
CHL	0.956	0.978
FL	0.946	0.972
GA	0.937	0.968
HC	0.931	0.965
Body	0.925	0.962
REL	0.924	0.961
LEL	0.918	0.958
AC	0.908	0.953
OCD	0.897	0.947
MFL	0.872	0.934
Liver	0.847	0.921
Kidneys	0.804	0.897
Lungs	0.800	0.894
RPFW	0.800	0.894
LPFW	0.781	0.884
ICD	0.743	0.862
Spleen	0.695	0.834
Adrenals	0.694	0.833
Thymus	0.679	0.824
PL	0.651	0.807
CC	0.572	0.756
HL	0.460	0.678
ID	0.406	0.637

4.2 Multiple Linear Regression Models

Across all variable selection methods for regression, outputs presenting models with non-significant variable coefficients were excluded (Enter and Remove). The Backward selection algorithm was discarded for presenting the same output as the Forward approach, while yielding a Durbin-Watson statistic further away from 2. Stepwise and Forward algorithms produced models with Durbin-Watson values of 1.961 and 1.958, respectively, and similar coefficients of determination ($R^2 \approx 0.953$). Both regressions share 5 retained variables, one exclusive variable each. Only statistically significant variable coefficients are present in either model (*p*-value $\leq$ 0.05).

Table 2. Standardized β-weights and variables selected by each regression algorithm method.

	Body	**FL**	**CHL**	**CRL**	**REL**	**Lungs**	**Adrenals**
Stepwise	0.402	0.310	0.266	-	0.157	-0.070	-0.087
Forward	0.384	0.384	-	0.199	0.163	-0.069	-0.083

4.3 Polynomial Regression Models

A collection of 115 single variable-based models for GA estimation were generated, 5 different degree polynomial regressions for each of the 23 independent variables. All models were retained after checking the statistical significance of each model's estimated parameters (p-value ≤ 0.05). R^2 values were stored for model comparison.

Table 3. R^2 values computed for all polynomial regressions. Polynomial degrees are represented by numbers 1 through 5, for each variable-derived model. Darker shades representing lower values. Table spliced due to size constraints.

	1	2	3	4	5
CHL	0.931	0.942	0.943	0.943	0.944
FL	0.927	0.940	0.942	0.945	0.945
Body	0.868	0.937	0.942	0.942	0.942
CRL	0.931	0.936	0.938	0.940	0.940
HL	0.410	0.917	0.930	0.934	0.936
HC	0.896	0.911	0.914	0.916	0.917
REL	0.893	0.902	0.904	0.907	0.907
LEL	0.885	0.891	0.895	0.896	0.896
Kidneys	0.734	0.876	0.877	0.881	0.881
CC	0.503	0.871	0.883	0.898	0.899
MFL	0.849	0.864	0.917	0.917	0.920
AC	0.840	0.840	0.852	0.853	0.857
Liver	0.759	0.840	0.842	0.843	0.843
OCD	0.834	0.835	0.854	0.857	0.860
Lungs	0.720	0.808	0.813	0.814	0.816
Spleen	0.623	0.791	0.833	0.847	0.849
RPFW	0.730	0.759	0.800	0.803	0.809
Thymus	0.608	0.756	0.816	0.820	0.820
LPFW	0.711	0.738	0.777	0.779	0.784
ICD	0.710	0.726	0.742	0.750	0.751
ID	0.363	0.715	0.722	0.777	0.787
Adrenals	0.589	0.681	0.689	0.691	0.692
PL	0.595	0.598	0.606	0.606	0.608

4.4 Comparison and Clustering

In terms of multiple linear regression, both previously selected models exhibited no statistically significant difference between them. In contrast, when either model was compared to any of the 115 polynomial regression models, a recurring p-value ≤ 0.05 was systematically observed.

By clustering models presenting no significant difference between other variable models, and creating different variable clusters based on statistical evidence for divergence, a goodness of fit hierarchy was established. CHL, CRL, and FL were the only single parameter-based regressions to be present in the top tier throughout all polynomial degrees. The hierarchical dissimilarities were most evident between 1st degree polynomial regressions and the remaining polynomial degree models.

Notably, body weight was placed alongside CHL, CRL, and FL as best GA estimators for any polynomial degree ≥ 2; HL was also classified in such terms for any polynomial degree ≥ 3. Generally, linear measurements outperformed weights in estimating GA. In addition, PCA and 1st degree polynomial clustering output the same variable hierarchy in terms of communality/loading values and R^2.

The following tables represent the outcome of polynomial regression clustering. Due to hierarchical ambiguity and/or redundancy, 3rd and 4th degree polynomial regression models were not included. Lower R^2 model clusters were also excluded due to size limitations.

Table 4. 1st degree polynomial regression goodness of fit clusters and ordered R^2. Darker shades representing lower values. Only top predictive variable clusters are present. Clusters are represented by boxes. Parameters in bold indicate cluster centre(s). For example, while AC and OCD models (first cluster centres) are statistically indistinguishable from MFL and one another, both have a significantly worse fit when compared to any other given model; MFL (second cluster centre) is statistically identical to Body, and both AC and OCD models, and significantly different from every other model.

0.931	CRL				
0.931	CHL				
0.927	FL				
0.896				HC	**HC**
0.893				REL	**REL**
0.885			LEL	**LEL**	LEL
0.868		Body	**Body**	Body	
0.849	MFL	**MFL**	MFL		
0.840	**AC**	AC			
0.834	**OCD**	OCD			

Table 5. 2nd degree polynomial regression goodness of fit clusters and ordered R^2. Darker shades representing lower values. Only top predictive variable clusters are present. Clusters are represented by boxes. Parameters in bold indicate cluster centre(s). Comparatively to the previous table, Body is now indistinguishable from any of the top 4 predictors.

0.942	CHL						
0.940	FL						
0.937	Body						
0.936	CRL						
0.917						HL	**HL**
0.911					HC	**HC**	HC
0.902				REL	**REL**	REL	
0.891			LEL	**LEL**	LEL		
0.876	Kidneys	Kidneys	**Kidneys**	Kidneys			
0.871	CC	**CC**	CC				
0.864	**MFL**	MFL	MFL				

Table 6. 5th degree polynomial regression goodness of fit clusters and ordered R^2. Darker shades representing lower values. Only top predictive variable clusters are present. Clusters are represented by boxes. Parameters in bold indicate cluster centre(s). Comparatively to the previous table, HL is now indistinguishable from any of the top 5 predictors.

0.945	FL				
0.944	CHL				
0.942	Body				
0.940	CRL				
0.936	HL				
0.920				MFL	**MFL**
0.917				HC	**HC**
0.907		REL	REL	**REL**	REL
0.899		CC	**CC**	CC	
0.896	LEL	**LEL**	LEL	LEL	
0.881	**Kidneys**	Kidneys			

5 Discussion and Final Remarks

In our case of 450 foetal autopsy cases, findings suggest that across all variables, CHL, CRL, and FL are the most appropriate candidate foetal parameters for GA estimation. For any degree of polynomial regression, these variables were always displayed within the significantly highest R^2 cluster. The same variables were also selected by multiple linear regression, exhibiting positive standardized β-weights ≥ 0.199 (ascendingly ordered CRL, CHL, and FL), and presented the highest PCA communality and loading values. Body weight, HC, HL, and ear length are also noteworthy candidate variables for either presenting high PCA communality and loading values, or having significantly meaningful β and/or R^2 values.

Accurately estimating foetal gestational age is essential for pregnancy management. As a further matter, GA estimation during autopsy procedures is key in assessing legal and criminal abortion cases. During these events, the estimation of gestational age depends on the foetal parameters used. Measurements of various foetal anthropometric features are frequently used for this purpose. Consistent with previously published work, CHL, CRL, and FL are found to be the most reliable sources of information for estimating developmental age. In cases where such measurements are impossible to obtain, other foetal features can be utilized (albeit less reliable) such as HL, HC, body weight, and ear length.

As our database evolves, and different foetal features are recorded, different studies can emerge. By analysing features such as cause of death and family background, in association with measurements and weights, machine learning algorithms can be executed to create a pathology prediction tool. This approach would be useful for early diagnosis of disease, aiding professionals and family in taking the appropriate action.

Acknowledgements This work was supported by FCT through funding of the LaSIGE Research Unit, ref. UID/CEC/00408/2013.

References

1. Hern WM. Correlation of fetal age and measurements between 10 and 26 weeks of gestation. Obstet Gynecol. 1984, 63(1): 26-32.
2. Gandhi D, Masand R, Purohit A. A simple method for assessment of gestational age in neonates using head circumference. Pediatrics. 2014, 3(5): 211-213.
3. Kumar GP, Kumar UK. Estimation of gestational age from hand and foot length. Med Sci Law. 1994, 34: 48-50.
4. Mercer BM, Sklar S, Shariatmadar A, Gillieson MS, D'Alton ME. Fetal foot length as a predictor of gestational age. Am J Obstet Gynecol. 1987, 156(2): 350-355.
5. Patil SS, Wasnik RN, Deokar RB. Estimation of gestational age using crown heel length and crown rump length in India. International J. of Healthcare & Biomedical Research. 2013, 2(1): 12-20.
6. Selbing A, Fjällbrant B. Accuracy of conceptual age estimation from fetal crown-rump length. J Clin Ultrasound. 1984, 12(6): 343-346.
7. Scheuer JL, MacLaughlin-Black S. Age estimation from the pars basilaris of the fetal juvenile occipital bone. Int J Osteoarchaeol. 1994, 4(4): 377-380.
8. Scheuer JL, Musgrave JH, Evans SP. The estimation of late fetal and perinatal age from limb bone length by linear and logarithmic regression. 1980, 7(3): 257-265.
9. Chikkannaiah P, Gosavi M. Accuracy of fetal measurements in estimation of gestational age. In J Pathol Oncol. 2016, 3(1): 11-13.
10. Gupta DP, Saxena DK, Gupta HP, Zeeshan Zaidi, Gupta RP. Fetal femur length in assessment of gestational age in thirds trimester in women of northern India (Lucknow, UP) and a comparative study with Western and other Asian countries. In J Clin Prac. 2013, 24(4): 372-375.
11. Archie JG, Collins JS, Lebel RR. Quantitative standards for fetal and neonatal autopsy. Am J Clin Pathol. 2006, 126(2): 256-265.
12. Sherwood RJ, Meindl RS, Robinson HB, May RL. Fetal age: methods of estimation and effects of pathology. Am J Phys Anthropo. 2000, 113(3): 305-315.
13. Andrews DT, Chen L, Wentzell PD, Hamilton DC. Comments on the relationship between principal components analysis and weighted linear regression for bivariate data sets. Chemometrics and Intelligent Laboratory Systems. 1996, 34(2): 231-244.
14. Nadaraya EA. On estimating regression. Theory of Probability & its Applications. 1964, 9(1): 141-142.
15. R Core Team. R: a language and environment for statistical computing, version 3.3.2. Vienna, Austria: R Foundation for Statistical Computing. 2016
16. Eaton JW, et al. GNU Octave version 3.0.1 manual: a high-level interactive language for numerical computations. CreateSpace Independent Publishing Platform. 2009.
17. Oliphant TE. Python for scientific computing. Computing in Science & Engineering. 2007, 9(3): 10-20.
18. Millman KJ, Aivazis M. Python for scientists and engineers. Computing in Science & Engineering. 2011, 13(2): 9-12.
19. Walt S, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation. Computing in Science & Engineering. 2011, 13(2): 22-30.
20. IBM Corp. IBM SPSS Statistics for Windows, version 24.0. Armonk, NY: IBM Corp. 2016.
21. Judd CM, McClelland GH, Ryan CS. Data analysis: a model comparison approach. Routledge. 2011.