

# Conditional Regression Model for Prediction of Anthropometric Variables

ERIK BROLIN\*<sup>†‡</sup>, LARS HANSON<sup>‡§</sup>, DAN HÖGBERG<sup>†</sup> and ROLAND ÖRTENGREN<sup>‡</sup>

<sup>†</sup> *Virtual Systems Research Centre, University of Skövde, Skövde, Sweden*

<sup>‡</sup> *Department of Product and Production Development, Chalmers University of Technology, Gothenburg, Sweden*

<sup>§</sup> *Industrial Development, Scania CV, Södertälje, Sweden*

---

## Abstract

In digital human modelling (DHM) systems consideration of anthropometry is central. Important functionality in DHM tools is the regression model, i.e. the possibility to predict a complete set of measurements based on a number of defined independent anthropometric variables. The accuracy of a regression model is measured by how well the model predicts dependent variables based on independent variables, i.e. known key anthropometric measurements. In literature, existing regression models often use stature and/or body weight as independent variables in so-called flat regressions models which can produce estimations with large errors when there are low correlations between the independent and dependent variables. This paper suggests a conditional regression model that utilise all known measurements as independent variables when predicting each unknown dependent variable. The conditional regression model is compared to a flat regression model, using stature and weight as independent variables, and a hierarchical regression model that uses geometric and statistical relationships between body measurements to create specific linear regression equations in a hierarchical structure. The accuracy of the models is assessed by evaluating the coefficient of determination,  $R^2$  and the root-mean-square deviation (RMSD). The results from the study show that using a conditional regression model that makes use of all known variables to predict the values of unknown measurements is advantageous compared to the flat and hierarchical regression models. Both the conditional linear regression model and the hierarchical regression model have the advantage that when more measurements are included the models will give a better prediction of the unknown measurements compared to the flat regression model based on stature and weight. A conditional linear regression model has the additional advantage that any measurement can be used as independent variable. This gives the possibility to only include measurements that have a direct connection to the design dimensions being sought. Utilising the conditional regression model would create digital manikins with enhanced accuracy that would produce more realistic and accurate simulations and evaluations when using DHM tools for the design of products and workplaces.

*Keywords: Anthropometry, Regression, Correlation, Multivariate, Prediction, Digital Human Modelling.*

---

## 1. Introduction

Digital human modelling (DHM) tools are used to reduce the need for physical tests and to facilitate proactive consideration of ergonomics in virtual product and production development processes (Chaffin et al. 2001; Duffy 2009). DHM tools provide and facilitate rapid simulations, visualisations and analyses in the design process when seeking feasible solutions on how the design can meet set ergonomics requirements. DHM software includes a digital human model, also called a manikin, i.e. a changeable digital version of a human. An important part of DHM systems is anthropometry, the study of human measurements,

and the functionality of creating human models based on a few predictive anthropometric measurements. The known predictive measurements, seen as independent variables, are used in a regression model to calculate a complete set of anthropometric measurements which are used to create digital human models that facilitates accurate ergonomics simulation and analyses. The number of independent key variables varies from case to case and should be chosen based on relevance to the design problem (Dainoff et al. 2004). Regression models can be seen as black boxes that use input, i.e. known key anthropometric measurements, to produce output, i.e. a complete set of anthropometric measurements (Figure 1).



**Figure 1** The regression model seen as a black box that uses input to produce output

The accuracy of a regression model should therefore be measured by how good the model predicts the unknown measurements, i.e. dependent variables, based on the known key anthropometric measurements, i.e. independent variables. In literature, concerning anthropometry, existing regression models often use stature and/or body weight as independent variables in linear regression equations (Drillis et al. 1966; Pheasant 1982; Gannon et al. 1998; Peacock et al. 2012). However, these so-called flat regressions models can make estimations with large errors when there are low correlations between the independent and dependent variables (You and Ryu 2005). To reduce this problem You and Ryu (2005) presented a hierarchical regression model that uses geometric and statistical relationships between body measurements to create specific linear regression equations in a hierarchical structure. Using a hierarchical regression model gives better estimates of predicted measurements if more measurements are known and used as input. Still, the hierarchical system requires measurements higher up in the hierarchy, i.e. stature and body weight, to be included in the analyses even if they do not necessarily have a direct connection to the design dimensions being sought (Bertilsson et al. 2011).

Using a conditional linear regression model that uses all known measurements to predict all unknown measurements would give better predictions and at the same time give the possibility to choose more freely which anthropometric measurements that should be used as input. It is possible to calculate the regression coefficients for a linear regression model through analysis of the correlation or covariance between known and unknown measurements (Johnson and Wichern 1992). This paper presents a conditional linear regression model and compares its predicted results with the results of a flat regression model based on stature and weight and a hierarchical regression model based on the method presented by You and Ryu (2005).

## 2. Materials and Methods

The conditional linear regression model analyses the covariance between the independent and dependent variables to calculate the regression coefficients. Based on the regression coefficients and the mean values, for both the independent and dependent variables, linear regression equations can

be constructed for each dependent variable. This multivariate statistical analysis is based on the assumption that anthropometric measurements can be approximated with a normal distribution, which holds true in most cases (Pheasant and Haslegrave 2006). However, the conditional linear regression model predicts the dependent variables with the smallest mean square error even if the normality assumption is not valid (Johnson and Wichern 1992). The statistical and mathematical analysis was done using MATLAB (MathWorks 2010) and ANSUR (Gordon et al. 1989) anthropometric data with measurements from 1774 males and 2208 females.

### 2.1. Mathematical procedure of the conditional regression model

A multivariate regression model uses  $k$  number of independent variables  $\mathbf{Z} = [Z_1, Z_2, \dots, Z_k]^T$  for the prediction of  $j$  number of dependent variables  $\mathbf{Y} = [Y_1, Y_2, \dots, Y_j]^T$  as

$$\mathbf{Y} = \boldsymbol{\beta}_o + \boldsymbol{\beta}\mathbf{Z} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_j \end{bmatrix} = \begin{bmatrix} \beta_{o_1} + \beta_{11} \cdot z_1 + \beta_{12} \cdot z_2 + \dots + \beta_{1k} \cdot z_k \\ \beta_{o_2} + \beta_{21} \cdot z_1 + \beta_{22} \cdot z_2 + \dots + \beta_{2k} \cdot z_k \\ \vdots \\ \beta_{o_j} + \beta_{j1} \cdot z_1 + \beta_{j2} \cdot z_2 + \dots + \beta_{jk} \cdot z_k \end{bmatrix}$$

When combining  $\mathbf{Y}$  and  $\mathbf{Z}$  the regression model gives a complete set of anthropometric measurements  $\mathbf{X} = [X_1, X_2, \dots, X_{j+k}]^T$  which is later used to describe joint centre positions and link lengths of a biomechanical model.

Suppose

$$\begin{bmatrix} \mathbf{Y} \\ \mathbf{Z} \end{bmatrix} = \begin{bmatrix} (j \times 1) \\ (k \times 1) \end{bmatrix} \text{ is distributed as } N_{j+k}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

with

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_Y \\ \boldsymbol{\mu}_Z \end{bmatrix} = \begin{bmatrix} (j \times 1) \\ (k \times 1) \end{bmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{YZ} \\ \boldsymbol{\Sigma}_{ZY} & \boldsymbol{\Sigma}_{ZZ} \end{bmatrix} = \begin{bmatrix} (j \times j) & (j \times k) \\ (k \times j) & (k \times k) \end{bmatrix}$$

The conditional expectation of  $\mathbf{Y}$ , given the fixed values  $\mathbf{Z}$  of the independent variables, is

$$E[\mathbf{Y} | Z_1, Z_2, \dots, Z_k] = \boldsymbol{\beta}_o + \boldsymbol{\beta}\mathbf{z} = \boldsymbol{\mu}_Y + \boldsymbol{\Sigma}_{YZ}\boldsymbol{\Sigma}_{ZZ}^{-1}(\mathbf{z} - \boldsymbol{\mu}_Z)$$

This conditional expected value, considered as a function of  $\mathbf{Z}$  is called the *multivariate regression* of the vector  $\mathbf{Y}$  on  $\mathbf{Z}$ . It is composed of  $j$  univariate regressions. The  $j \times k$  matrix

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{YZ}\boldsymbol{\Sigma}_{ZZ}^{-1}$$

is called the matrix of *regression coefficients* and the  $j \times 1$  vector

$$\beta_o = \mu_Y - \Sigma_{YZ} \Sigma_{ZZ}^{-1} \mu_Z$$

is the vector containing the intersection point for each regression equation.

## 2.2. Description of comparison procedure

The described conditional regression model was compared to a flat regression model based on stature and weight and a hierarchical regression model based on the method presented by You and Ryu (2005). In the analyses gender was treated separately by creating specific regression equations for each gender for the flat and hierarchical regression models and letting the conditional regression model analyse both female and male data. 56 anthropometric measurements (Table 1) were included in the analysis and four comparative tests were done where the number of independent variables varied for each test. The first test was done with stature and weight as independent variables which are the measurements that are necessary in the flat and hierarchical regression models. The second and third test was done using the first 7 and 17 measurements respectively according to Table 1. These measurements were chosen as they could be found high up in the hierarchal model described by You and Ryu (2005). The last and final test was done using the last three measurements in Table 1, hip breadth (sitting), popliteal height and radiale-styilion length. These measurements are found further down in the hierarchy of the hierarchical model, but could still be interesting to use, for example in the design of an office chair. The last test was not possible to perform with the flat and hierarchical model, since these models require stature and weight as independent variables, but was useful to show the capability of the conditional regression model. The three regression models were compared by assessing the coefficient of determination,  $R^2$ , as

$$R^2 = \frac{\sum_{j=1}^n (\hat{y}_j - \bar{y})^2}{\sum_{j=1}^n (y_j - \bar{y})^2}$$

and the root-mean-square deviation (RMSD) as

$$RMSD = \sqrt{\frac{\sum_{j=1}^n (y_j - \hat{y}_j)^2}{n}}$$

where  $y_j$  is the measured value,  $\bar{y}$  the mean value,  $\hat{y}_j$  the predicted value and  $n$  the number of measured individuals (1774 males and 2208 females).

**Table 1** The 56 anthropometric measurements included in the evaluation of regression models

#	Anthropometric measurement
1	Stature
2	Weight
3	Acromial height
4	Knee height at midpatella
5	Trochanterion height
6	Thumb-tip reach
7	Waist circumference at omphalion
8	Buttock circumference
9	Chest circumference
10	Elbow circumference
11	Forearm-hand length
12	Functional leg length
13	Hand circumference at metacarpale
14	Hand length
15	Head circumference
16	Thigh circumference, proximal
17	Wrist circumference, stylium
18	Ankle circumference
19	Axilla height
20	Arm circumference at axillar
21	Foot circumference
22	Biacromial breadth
23	Bideltoid breadth
24	Buttock depth
25	Buttock-knee length
26	Buttock-popliteal length
27	Calf circumference
28	Cervicale height
29	Chest breadth
30	Chest depth
31	Crotch height
32	Eye height (sitting)
33	Foot breadth
34	Foot length
35	Forearm circumference, flexed
36	Gluteal furrow height
37	Hand breadth at metacarpale
38	Head breadth
39	Head length
40	Heel breadth
41	Hip breadth
42	Interpupillary distance
43	Knee circumference
44	Knee height (sitting)
45	Lateral malleolus height
46	Neck circumference over larynx
47	Shoulder-elbow length
48	Sitting height
49	Thigh clearance
50	Waist breadth at omphalion
51	Waist depth at omphalion
52	Waist height at omphalion
53	Wrist to centre-of-grip length
54	Hip breadth (sitting)
55	Popliteal height
56	Radiale-Stylium length

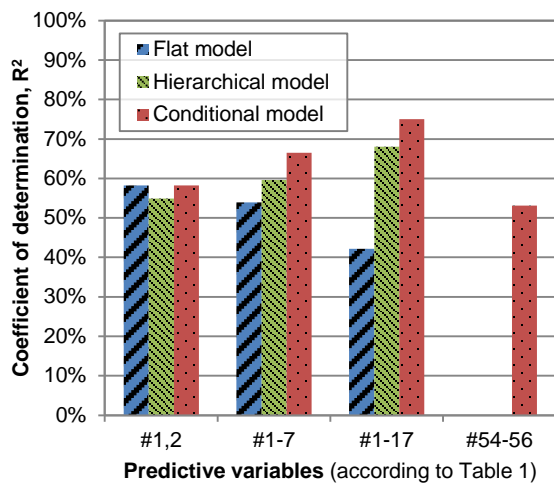
## 3. Results

In the regression models gender was treated separately and the coefficient of determination and the root-mean-square deviation was calculated for each dependent variable for each test. However, only the combined average results, for both genders and the dependent variables for each test, are presented in the following text and figures (Table 2, Figure 2 and Figure 3).

**Table 2** Average  $R^2$  value and the root-mean-square deviation for the dependent variables for each test

Predictive variables	Regression model	$R^2$ value	RMSD
Test 1: #1,2	Flat	58.2%	13.88
	Hierarchical	54.9%	14.27
	Conditional	58.2%	13.88
Test 2: #1-7	Flat	53.9%	15.29
	Hierarchical	59.6%	11.32
	Conditional	66.5%	10.07
Test 3: #1-17	Flat	42.1%	19.22
	Hierarchical	68.0%	9.60
	Conditional	75.0%	8.02
Test 4: #54-56	Flat	N/A	N/A
	Hierarchical	N/A	N/A
	Conditional	53.1%	16.16

In test 1, when stature and weight was used as independent variables the resulting  $R^2$  value and root-mean-square deviation were approximately the same for all three regression models. However when the number of independent variables increases the accuracy of the flat regression model decreases compared to the hierarchical and conditional regression models. In test 2, when 7 measurements were used as independent variables, the conditional model had an average  $R^2$  value that was 23.3% higher than that of the flat model and 11.6% higher than that of the hierarchical model. Analysis of root-mean-square deviation showed a decrease for the conditional model by 34.2% compared to the flat model and 11.1% compared to the hierarchical model.

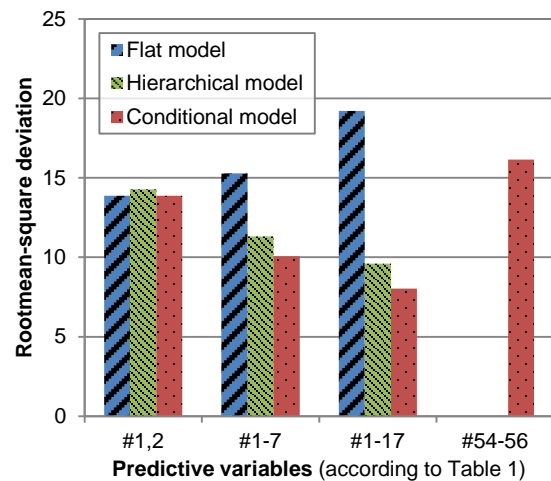


**Figure 2** Graph illustrating the evaluation of the  $R^2$  value for the dependent variables based on the results from the four different tests (*Higher value indicates higher accuracy*)

In test 3, when 17 measurements were used as independent variables, the conditional model had an average  $R^2$  value that was 78.0% higher than that of the flat model and 10.2% higher than that of the hierarchical model. Analysis of root-mean-square deviation showed a decrease for the conditional

model by 58.2% compared to the flat model and 16.4% compared to the hierarchical model.

In test 4, when the last three measurements in Table 1 were used as independent variables, results could only be attained from the conditional regression model as it can use any variables as input. However, the results from test 4 show a decrease in accuracy of predicting dependent variables. Compared to test 1, with stature and weight as independent variables, test 4 shows a decrease in  $R^2$  value by 8.70% and an increase in root-mean-square error of 16.4%.



**Figure 3** Graph illustrating the evaluation of the root-mean-square deviation for the dependent variables based on the results from the four different tests (*Lower value indicates higher accuracy*)

In total the conditional regression model shows the highest accuracy when predicting unknown variables. For the first three tests the conditional model had, on average, an accuracy that was 31.7% higher than that of the flat model and 9.3% higher than that of the hierarchical model (depending on if the coefficient of determination or the root-mean-square deviation was assessed). In test 4 the conditional model was the only model that could produce any results.

#### 4. Discussion

The results from the study shows that using a conditional regression model that makes use of all known variables to predict the values of unknown measurements is advantageous compared to the flat and hierarchical regression models. Both the hierarchical regression model and the conditional linear regression model have the advantage that when more measurements are included the models will give a better prediction of the unknown measurements compared to the flat regression model based on two variables, stature and weight. A conditional linear regression model has the additional advantage that any measurement can be

used as independent variable. This gives the possibility to only include measurements that have a direct connection to the design dimensions being sought. For example, in the case of creating multidimensional boundary cases it is of interest to reduce the number of anthropometric measurements that are used as input, i.e. it is advantageous to not always be forced to include stature and weight (Brolin et al. 2012). In other cases, stature and weight might not be of interest to include in the analysis, e.g. when designing a shoe, helmet or a hand control. However, using a DHM tool to evaluate the design of the product could still be of interest and thus requiring the functionality of the conditional regression model.

Sorting the mean vector and the covariance matrix by independent and dependent variables and by using matrix algebra as described in section 2.1 eases the process of defining regression equations for the dependent variables. The comparison of the different regression models was done using ANSUR anthropometric data (Gordon et al. 1989). This data is not representative for any civilian population today (since it was measured 20 years ago on army personnel) but considered relevant here in that it covers large data set of both measurements and individuals. The presented conditional regression model could easily be applied to any anthropometric data, which can be approximated by a normal distribution and for which the mean value and covariance matrix are known or can be calculated. Even though the described method makes good prediction of dependent variables the diversity variance of anthropometric data is not considered. A manikin based on a number of specific measurements will, by using the presented model, always look the same. This is not the case in human populations, e.g. persons of a specific stature will most likely have different weights and proportions. Parkinson and Reed (2010) proposes a model for creating virtual user populations which also incorporates a stochastic component retaining relevant variance of the anthropometric data. The presented conditional model would be possible to extend to calculate and incorporate the variance for each dependent variable based on the independent variables. This would give digital human models with anthropometry that better resembles the variance and diversity that exist within human populations. This would in turn produce more realistic and accurate simulations and evaluations and thus give better assistance to engineers and designers using DHM tools when developing products and workplaces.

## 5. Conclusion

Results from the study shows that the conditional model produces more accurate predictions compared to a flat regression model based on stature and weight, and also to a hierarchical regression model that uses geometric and statistical relationships between body measurements to make predictions. Utilising the conditional regression model would create digital manikins with enhanced accuracy that would produce more realistic and accurate simulations and evaluations when using DHM tools for the design of products and workplaces.

## Acknowledgement

This work has been made possible with the support from the Swedish Foundation for Strategic Research/ProViking and by the participating organisations. This support is gratefully acknowledged.

## References

- Bertilsson, E, Hanson, L, Högberg, D, Rhén, I-M, 2011. Creation of the IMMA manikin with consideration of anthropometric diversity. Proceedings of the 21st International Conference on Production Research (ICPR). Fraunhofer Verlag.
- Brolin, E, Högberg, D, Hanson, L, 2012. Description of boundary case methodology for anthropometric diversity consideration. *International Journal of Human Factors Modelling and Simulation*, 3, 204-223.
- Chaffin, D B, Thompson, D, Nelson, C, Ianni, J D, Punte, P A, Bowman, D, 2001. Digital human modeling for vehicle and workplace design, Warrendale, PA, Society of Automotive Engineers.
- Dainoff, M, Gordon, C, Robinette, K M, Strauss, M, 2004. Guidelines for using anthropometric data in product design. HFES Institute Best Practices Series. Santa Monica: Human Factors and Ergonomics Society.
- Drillis, R, Contini, R, Bluestein, M, 1966. Body Segment Parameters. Technical Report No. 1166.03. NTIS No. PB 174 945. New York: New York University, School of Engineering and Science, Research Division.
- Duffy, V G, 2009. Handbook of Digital Human Modeling, Boca Raton, CRC Press.
- Gannon, A J, Moroney, W F, Biers, D W, 1998. The validity of anthropometric predictions derived from proportional multipliers of stature. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. SAGE Publications.
- Gordon, C C, Churchill, T, Clauser, C E, Bradtmiller, B, Mcconville, J T, Tebbetts, I, Walker, R A, 1989. 1988 Anthropometric Survey of US Army Personnel: Methods and Summary Statistics. Technical Report Natick/TR-89-044.

- Natick, MA.: U.S. Army Natick Research, Development and Engineering Center.
- Johnson, R A, Wichern, D W, 1992. Applied multivariate statistical analysis, Englewood Cliffs, NJ, Prentice-Hall, Inc.
- Mathworks, 2010. MATLAB version 7.10.0.499 (R2010a). Natick, Massachusetts: The MathWorks Inc.
- Parkinson, M B, Reed, M P, 2010. Creating virtual user populations by analysis of anthropometric data. *International Journal of Industrial Ergonomics*, 40, 106-111.
- Peacock, J B, Manoharan, A, Tong, X, Ping, C Y, Ping, L W, Ding, F, Chuan, T K, Hartono, M, Stella, N Y, 2012. Drillis and Contini Revisited. In: REBELO, F. & SOARES, M. M. (eds.) *Advances in Usability Evaluation, Part II*. Boca Raton: CRC Press.
- Pheasant, S, 1982. A technique for estimating anthropometric data from the parameters of the distribution of stature. *Ergonomics*, 25, 981-992.
- Pheasant, S, Haslegrave, C M, 2006. *Bodyspace: Anthropometry, Ergonomics and the Design of Work*, Boca Raton, CRC Press.
- You, H, Ryu, T, 2005. Development of a hierarchical estimation method for anthropometric variables. *International Journal of Industrial Ergonomics*, 35, 331-343.