Is Statistics Making Human Factors Researchers Lose Touch with Reality?

Goonetilleke, Ravindra S.

Human Performance Laboratory / Department of Industrial Engineering and Logistics Management / Hong Kong University of Science and Technology / Clear Water Bay / Hong Kong

E-mail: ravindra@ust.hk

ABSTRACT

Statistical results are basic requirements to show significant effects. Discussing insignificant findings is considered as a stretch in relation to any experiment. Significance depends on many factors amongst which the sample size and the range of the dependent variables play an important role. Thus it is important to consider these aspects as well. In this paper, I intend to show that even models with low explained variances can have value depending on the accuracy and the precision that is required. Using foot anthropometric data and regression analysis, I show that the simpler forms of regression equations can be more useful, even though, at the expense of an increase in error variances, especially in relation to dimensions such as foot length, arch length and foot width.

Keywords

Regression, Foot anthropometry, Footwear design, Foot length, Foot width, Arch Length

INTRODUCTION

With the availability of high computing power and other sophisticated devices, researchers as well as layman attempt to utilize these resources as much as possible. The days of printing charts and sticking them on walls is no more. When the author was a graduate student, pages of charts were printed on a dot-matrix printer and they were pasted wherever they were visible throughout the day. Today, such practices would be looked down upon as an environmentally-unfriendly act. We tend to ignore visible patterns and try to make sense and use regression and other statistical models as much as possible. A simple linear regression model is expressed as:

$$Y = \beta_0 + \beta_1 X$$

It is also known as a least-squares fit or a linear regression model. The expression ($\beta_0 + \beta_1 X$) is known as the deterministic component of the model. Strictly speaking any point in the X, Y space can be expressed in terms of the probabilistic model, which is:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

In here ε is random error component. The assumptions are that for each X, ε has a normal probability distribution with mean 0 and a constant variance σ^2. It is also assumed that for every possible pair of observations Y_i and Y_j, the associated random errors ε_i and ε_j are independent. From a statistical viewpoint, independence is important in many models. However, in practice, we would like to be able to predict Y for any given X value. If the random error at one setting is known, and if the errors were dependent, we would be able to obtain an accurate estimate of Y. Unfortunately, this is not the case. The goodness of fit of the line is determined or expressed as the coefficient of determination (R^2) with its values between 0 and 1. The R^2 value is an indication of the variance explained by the model. Values of at least 0.7 are used in practice in order to accept a relationship between X and Y [1]. In this paper, I will show some examples of regression lines and attempt to show how they can rejected even though they can be useful in predicting the magnitude of the corresponding dependent variable.

METHODOLOGY

The staff and students in the Human Performance Laboratory have performed many experiments in relation to human feet. Hence I would use one subset of data, part of what has been reported in Witana et al. [3] for this analysis. Twenty-five males and twenty-five females participated in this experiment. The weight, height, foot length, arch length, foot width, the dorsal foot height at 50% foot length (mid-foot height) and many other dimensions were measured by two operators twice. The reliability of the measurements has been reported in Witana et al. [3]. Here, I use the mean values of the four readings for the statistical analyses that were performed using Minitab, SAS and Excel.

RESULTS and DISCUSSION

The descriptive statistics are shown in Table 1.

Table 1. Descriptive Statistics of 50 subjects

	Mean	SD	Min	Max
Age	21.5	1.2	19	24
Height (cm)	165.0	8.2	148.2	189.7
Weight (kg)	55.9	9.4	39.1	79.1
Foot Length (cm) (FL)	24.7	1.5	21.4	29.0
Arch Length (cm)	18.0	1.0	16.05	20.65
Foot Width (cm)	9.5	0.6	8.25	10.7
Height at 50% FL (cm)	5.9	0.5	4.93	7.0

It is no doubt good to be able to predict foot dimensions from the weight and height of a person. In the forensic sciences there is a need to be able to estimate the weight and the height of a person from a foot print. Hence, such models have value. Consider the case of arch length and foot length. The arch length is important in the design of the flex groove of a shoe. Rather than measure it, if it can be predicted from a person's foot length, it would help manufacturers design more flexible and functional shoes.

The foot length and arch length measures are normally distributed (Foot length: Anderson-Darling statistic = 0.389, p =0.373; Kolmogorov-Smirnov statistic= 0.097, p > 0.15. Arch Length: Anderson-Darling statistic = 0.343, p = 0.476; Kolmogorov-Smirnov statistic= 0.075, p > 0.15). Hence it would be possible to investigate the relationship between foot lengths and arch lengths.

The relationship between arch length and foot length when all the available subject data are used is as follows:

Arch length = 1.61 + 0.666 Foot Length(cm) R^2 = 0.931 R^2(adj) = 0.929 (1)

where the adjusted R^2 is a reduced value of R^2 to make an estimate of the R^2 of the population and calculated as:
Adjusted $R^2 = 1 - (1- R^2)(N-1)/(N-k-1)$
where k is the number of independent variables [2]. k=1 for the above case. With large values of N, R^2 will be close to the Adjusted R^2. The scatterplot and regression line are shown in Figure 1 and the residuals (ε) from the prediction are depicted in Figure 2. The residuals are normally distributed with a maximum residual of (-)0.69 cm.

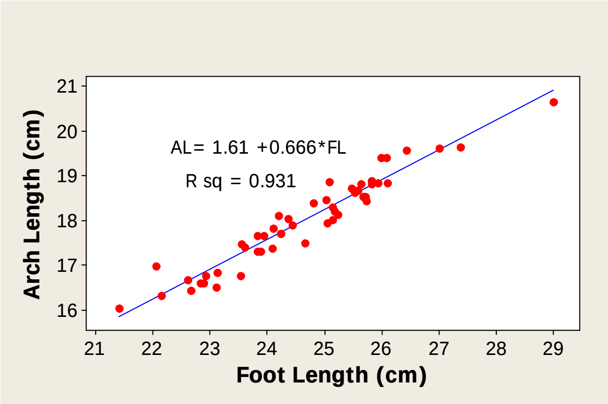

Figure 1. Scatter plot of arch length versus foot length with line of best fit.

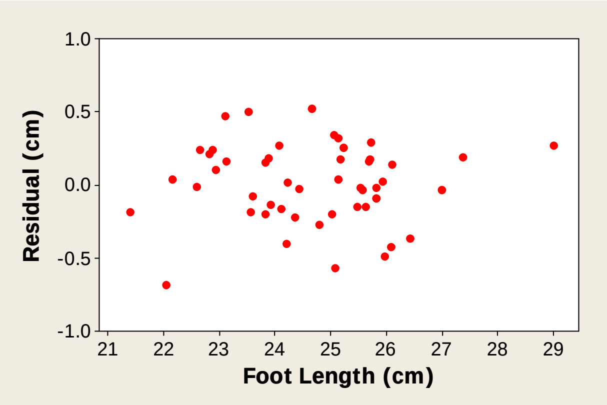

Figure 2. Residual plot of arch length using equation (1)

The corresponding analysis of variance from the regression is shown in Table 2. The value R^2 value is calculated as $SS_{regression}/SS_{total}$. The null hypothesis is that R^2= 0. Table 2 shows that $p < 0.05$ and hence R^2 is significant and is greater than zero. The second part of the equation investigates the effect of the two coefficients in the regression equation, namely the intercept (=1.61) and the slope (=0.666). The p value of less than .05 indicates that there is a significant intercept and slope and that it is unlikely to be zero in the population.

Table 2. Analysis of Variance

Source	**DF**	**SS**	**MS**	**F**	**P**
Regression	1	47.877	47.877	646.66	0.000
Residual Error	48	3.554	0.074		
Total	49	51.430			

Predictor	**Coefficient**	**SE Coef.**	**T**	**P**
Constant	1.61	0.6469	2.48	0.017
Foot Length(cm)	0.666	0.02619	25.43	0.000

Minitab shows that two of the observations result in a large standardized residual (residual = 0.69; standardized residual= 2.64 and residual = 0.725; standardized residual=2.13) and one observation whose foot length value (=29 and shown to the right in the plot of arch length vs. foot length) gives it large leverage (residual = -

0.266; standardized residual=-1.09). The regression was re-run eliminating these three observations. The resulting equation is:

Arch length = 0.784 + 0.698 Foot Length(cm) $R^2 = 0.940$ $R^2(adj) = 0.938$ (2)

The analysis of variance showed that the regression and the slope are significant, but the intercept (constant term) is not significant (in other words it is likely to be zero in the population). The statistical analysis makes perfect sense as an arch length of 0.784 cm is not meaningful if the foot length is zero unless there is a measurement bias. Generally, arch length is measured when foot length is measured and hence there is no reason for a constant term. So a regression line with a zero intercept ought to be more appropriate and will be presented later. Minitab shows that two of the observations (residual = -0.4955; standardized residual=-2.11 and residual = 0.4752; standardized residual=2.03) results in a large standardized residual and one observation gives it large leverage (residual = 0.3203; standardized residual=1.45). Hence these three observations were eliminated and the regression re-run. The resulting equation is:

Arch length = 0.632 + 0.704 Foot Length(cm) $R^2 = 0.946$ $R^2(adj) = 0.944$ (3)

Even at this stage, one observation shows a large residual (residual=0.4492 standardized residual=2.09). Hence that was eliminated as well. The resulting equation is:

Arch length = 0.547 + 0.707 Foot Length(cm) $R^2 = 0.951$ $R^2(adj) = 0.950$ (4)

As in the previous step, the analysis of variance showed that the regression and the slope are significant, but the constant is not significant. Now two more observations have large standardized residuals (residual = 0.4099; standardized residual=2.02 and residual = -0.4114; standardized residual=2.01). The above process continues with small increases in R^2. It also appears that the elimination of the outliers, help reduce the magnitude of the intercept. Thus, the non-significant intercept suggests an alternative way to model the relationship between arch length and foot length with zero intercept as follows:

Arch length = 0.7308 Foot Length(cm) (5)

Minitab does not provide a R^2 value with a zero intercept. However, Excel indicates that $R^2 = 0.922$. The residual plots with equations (1) and (5) are shown in Figures 2 and 4. The distribution of the residuals is in Table 3, which shows that 92% of the residuals are less than 5mm. Even though the residuals are normally distributed, the scatterplots show some potential outliers (Figures 2 and 4). Only the zero intercept condition will be considered here. Eliminating the observation with the highest residual value, and the two values of foot length corresponding to the highest and lowest (extremes), the regression equation becomes:

Arch length = 0.730 Foot Length(cm) (6)

Considering equations (2) and (6), it is quite evident that equation (2) has a higher R^2 value with almost identical residual values. It is clear that equation (6), with a zero intercept, which states that arch length is 73% of foot length, is much easier to use and apply. The resulting error from the use of this equation is a maximum of approximately 0.57 cm whereas the use of an equation with an intercept (i.e., equation 2) will result in a maximum error of 0.4955 mm. The amount of explained variance foregone is reflected in the very slight increase in the residual error. If the change in error from using equation (2) and (6) is not of importance, it would be wiser to use equation (6). On the flip side, if the intercept of 0.784 cm in equation (2) can be explained, it may be reasonable to use that equation.

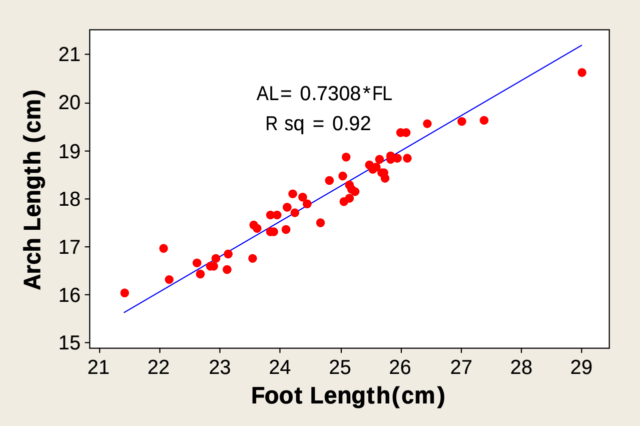

Figure 3. The relationship of Arch length with foot length (equation 5).

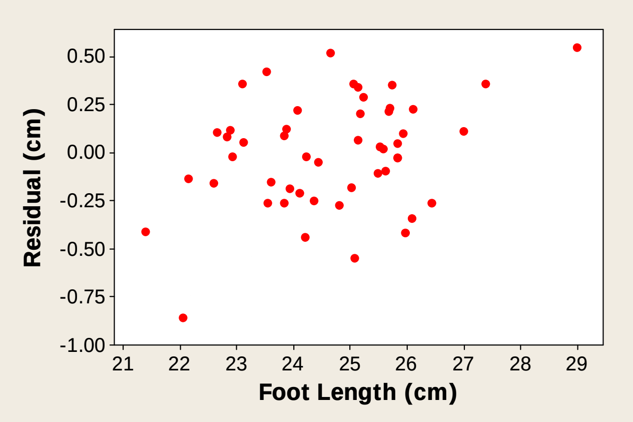

Figure 4. Residual plot with equation (5)

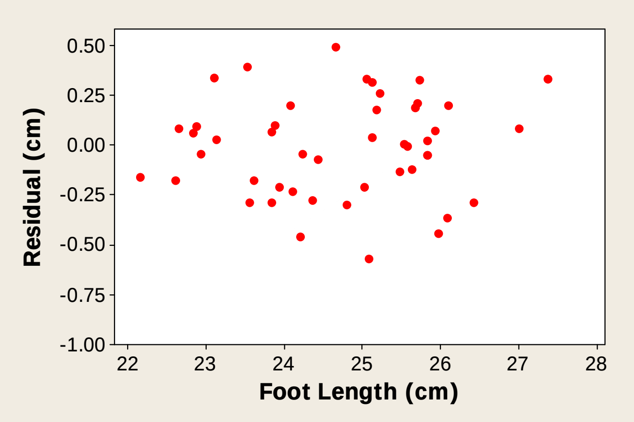

Figure 5. Residual plot with equation (6)

Table 3. The residual distribution using equation (5)

Residual (ε) mm	Number of observations	Percentage
ε >5	4	8
4<ε≤5	4	8
3<ε≤4	6	12
2<ε≤3	11	22
1<ε≤2	10	20
ε≤1	15	30
Total	**50**	**100**

The Foot Length vs. Height, Foot width vs. Weight and Mid-foot Height vs. Height regressions and the corresponding residuals are shown in Figures 6 to 11.

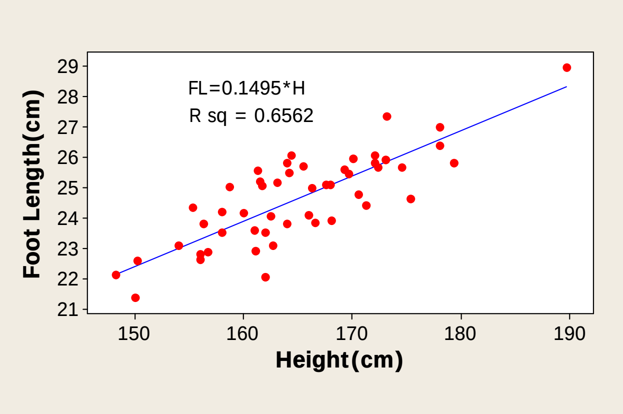

Figure 6. The relationship of foot length with height.

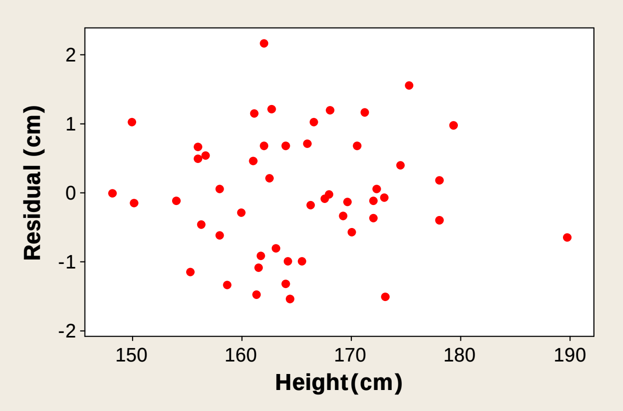

Figure 7. The residual plot of foot length vs. height with FL=0.1495H

Figures 1 and 3 show regression lines with $R^2 \geq 0.9$. From a statistical viewpoint this can be considered to be a reasonable fit. However, Figures 6, 8, and 10 do not show such a high value of R^2. The R^2 values are 0.66, 0.41 and 0.40. These may be considered regression lines having poor fit and thus one may conclude that the prediction from the model is inadequate. However, consider the case of Figure 6. The regression line shows that:

$$\text{Foot Length} = 0.1495\ (\text{Height}) \qquad (7)$$

If this equation is used, the residual values (random error) can be calculated for each data point. The maximum residual in this case is 2.2 cm and the minimum is -1.5 cm. The residual (Figure 7) appears to have a random distribution. Even though the maximum residuals may be considered large, Figure 7 shows that a large proportion (70%) of the data points has a residual of less than 1 cm. For a variable such as foot length, would it be inappropriate to be able to predict foot length from height with a 70% probability that the error is less than 1 cm? For shoe manufacture, especially with clearance at the toe areas, the prediction accuracy ought to be sufficient. An alternative way to derive proportions is to use the confidence intervals.

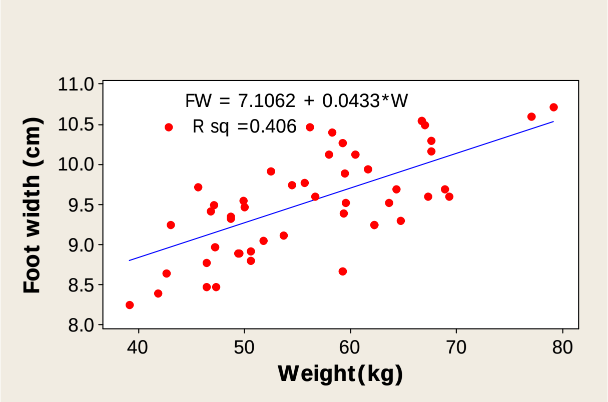

Figure 8. The relationship of foot width with weight.

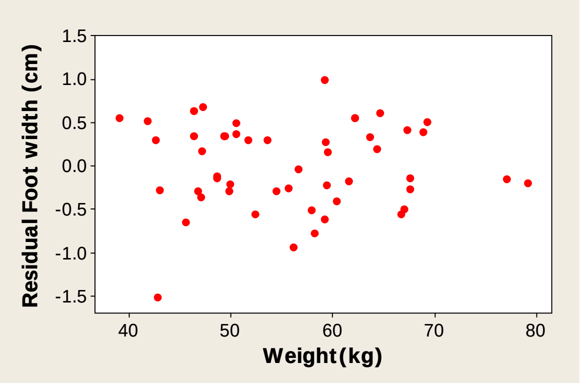

Figure 9. The residual plot of foot width vs. weight.

The range of foot width across the experimental subjects is 8.25 cm to 10.7 (Table 1). Mid-foot height shows a relatively smaller range (4.93 to 7 cm) as well (Table 1). The regression equation for foot width is as follows:

$$\text{Foot width} = 7.11 + 0.0433\ \text{Weight(kg)} \quad R^2 = 40.6\% \quad R^2\,(\text{adj}) = 39.3\% \qquad (8)$$

The above equation can predict foot width to within 5 mm 64% of the time. The analysis of variance is shown in Table 4. If 5 mm is tolerable, then the equation with such a low R^2 ought to be useful.

Table 4. Analysis of Variance related the regression of Foot width and weight

Source	**DF**	**SS**	**MS**	**F**	**P**
Regression	1	8.1365	8.1365	32.74	0.000
Residual Error	48	11.9277	0.2485		
Total	49	20.0642			

Predictor	**Coefficient**	**SE Coef.**	**T**	**P**
Constant	7.1062	0.4292	16.56	0.000
Weight (kg)	0.043339	0.007574	5.72	0.000

Figures 10 and 11 show the relationship and the residuals associated with modeling mid-foot height from the height of a person. The analysis of the residuals reveal that mid-foot height can be predicted to within 0.5 cm 80% of the time.

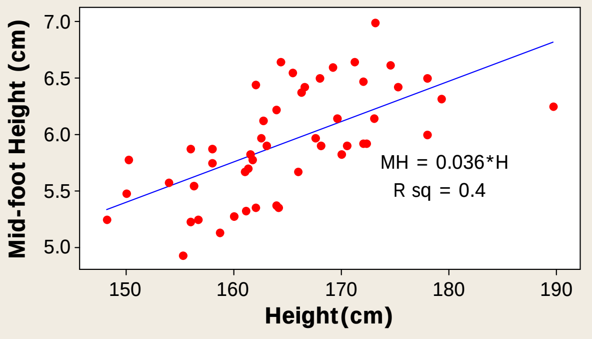

Figure 10. The relationship of mid-foot height with height.

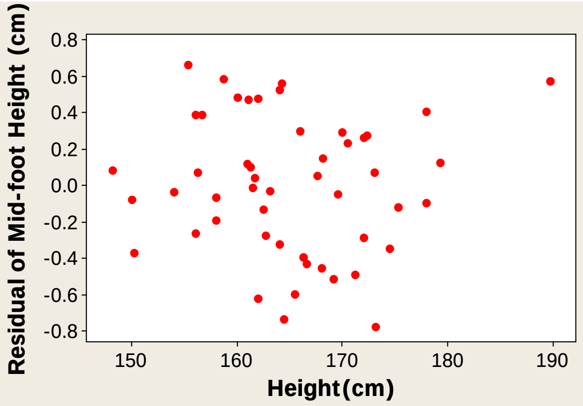

Figure 11. The residual plot of mid-foot height with height.

CONCLUSIONS

As can be seen from all the regression analyses and their interpretations, the appropriateness of the prediction models depends to a large extent on the accuracy that is required. Even though the explained variance of a model is relatively low, the predicted value may still be useful depending on the use of that prediction. Thus, it is important to know the basic needs prior to throwing-out any potential relationship even though the R^2 value may be relatively low.

ACKNOWLEDGEMENTS

The author would like to thank the Research Grants Council of Hong Kong for funding this study under grant HKUST 613406.

REFERENCES

[1] Department of Defense, 1995, Parametric Cost Estimating Handbook, http://cost.jsc.nasa.gov/PCEHHTML/pceh.htm, last accessed on 10 February 2008.

[2] Miles, J. and Shevlin, M., 2001, Applying regression and Correlation, Sage Publications, London.

[3] Witana, C.P., Xiong, S., Zhao, J. and Goonetilleke, R. S., 2006, Foot measurements from three-dimensional scans: A comparison and evaluation of different methods. International Journal of Industrial Ergonomics September, Vol. 36(9), pp. 789-807.