Exploring Relationships between Body Dimensions in Physically Active Individuals

CFAS440 - Generalised Linear Models I

Harry Baines
35315878
h.baines3@lancaster.ac.uk

Owen Dwyer
35283985
o.dwyer@lancaster.ac.uk

Abstract

Multiple linear regression was carried out to investigate the relationship between a person's weight and a range of body measurements for physically active individuals. We fit a regression model to accurately predict a person's weight based on multiple explanatory variables. We performed stepwise model selection to identify the most parsimonious model and identify interaction effects between the variables. Regression diagnostics were used to evaluate the model assumptions - the data met the assumptions of homogeneity of variance and linearity, and the residuals were approximately normally distributed. Through application of the Box-Cox transformation, we found that modifying the weight by the fourth root further improved model fit.

1 Introduction

1.1 Background and Motivation

The HP Study was a study which aimed to investigate the relationships between body build, weight, and girths in a group of physically active young men and women [1]. Most of the participants in the study were within the normal weight range. The dataset associated with the study contains body girth and skeletal diameter measurements, as well as age, height, weight and gender for 507 physically active individuals (247 men and 260 women). In the study it was hypothesised that the height and various skeletal variables predict scale weight significantly better than just height measurements.

1.2 Aims

This paper aims to discover the key skeletal and girth measurements which influence the weight of an individual. This will be achieved by fitting the best multiple linear regression model to the body measurements data, such that the weight of an individual can be accurately predicted using a regression equation for new and unseen data. We also aim to discover interaction effects for our explanatory variables which influence the value of a person's weight.

1.3 Data Description

The data are described by 25 features, as summarised in Table 1. There were no missing values. The measurement sites and techniques were chosen according to the procedure recommended by Behnke and Wilmore [2]. Nine of these are skeletal diameter measurements taken using an anthropometer. Soft tissue was compressed to ensure that the measurements were taken "bone to bone". A further twelve features represent girth or circumference measurements. In contrast to the skeletal measurements, all but three of these tend to change over a human's lifespan. The final four features age, height, weight and gender were also recorded for all individuals. All variables except for gender are quantitative and continuous in nature.

Measurements

Skeletal (cm)	Girth (cm)	Other
Biacromial *(shoulder breadth)*	Wrist*	Age (years)
Biiliac *(hips)*	Knee*	Height (cm)
Bitrochanteric *(thighs)*	Ankle*	Weight (kg)
Chest depth	Shoulder	Gender [M/F]
Chest diameter	Chest	
Elbow	Waist	
Wrist	Navel	
Knee	Hip	
Ankle	Thigh	
	Bicep	
	Forearm	
	Calf	

* *generally constant over lifetime*

Table 1: Dataset features

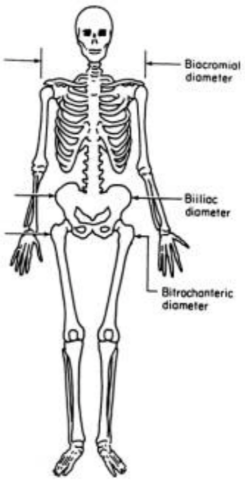

Figure 1: Selected measurement locations [1]

It is important to note that the sampling was not random or stratified. Participants were described as all being physically active, and were recruited from locations including a U.S. Navy school and various health and fitness clubs, suggesting that participants in this sample may be significantly physically fitter than the general population. This becomes apparent when the data is analysed. For example, it is widely established that on average, men exceed women in all of the explanatory measurements in the dataset with the exception of hip and thigh girth [3]. However, the dataset's authors observe that in this dataset, women only exceed men on average in thigh girth [1]. It is therefore important to consider these limitations and exercise caution when attempting to make inferences about the general population based on this data alone.

1.4 Report Outline

Section 2 describes our exploratory analysis of the data, including plots of response variables against chosen explanatory variables and histograms to understand the data distribution, calculations of correlations between variables, and a brief summary of descriptive statistics of the data. In Section 3 we implement a main effects model and analyse the output following backwards selection. We also identify interaction effects between the variables and provide graphical plots of the results. In Section 4 we use diagnostic tests to ensure the assumptions which were made in the modelling section all hold, and make adjustments accordingly. Finally, Section 5 concludes with a summary of our discoveries, highlighting the significance of what was found, and suggestions for further analysis.

2 Exploratory Analysis

In this section we will analyse the data without making any major assumptions. Our analysis will involve investigating relationships between the variables, identifying any potential outliers, using graphical methods to understand the distribution of the data and to display correlations among the variables, and non-graphical methods to calculate summary statistics for the data.

2.1 Graphical Analysis

The box-plot in Figure 2 shows roughly symmetrical distributions of weight for each gender. We can also see a higher median weight for males compared to females. The box-plot also highlights potential outliers for both males and females, although we do not consider these values to be extreme enough to disregard in our analysis.

The histogram in Figure 3 shows a positive skew of ages for both males and females. We notice most of the participants were aged 20-30 in both gender categories, with ages at the upper end of nearly 70 years old.

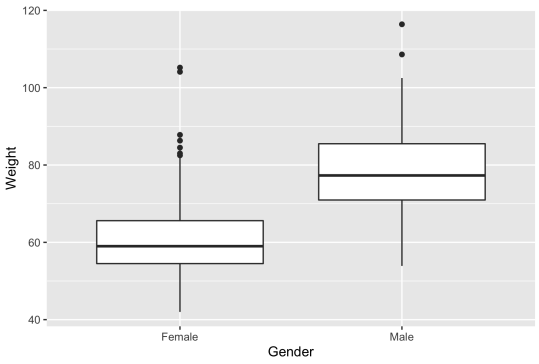

Figure 2: Box-plot of Weight against Gender.

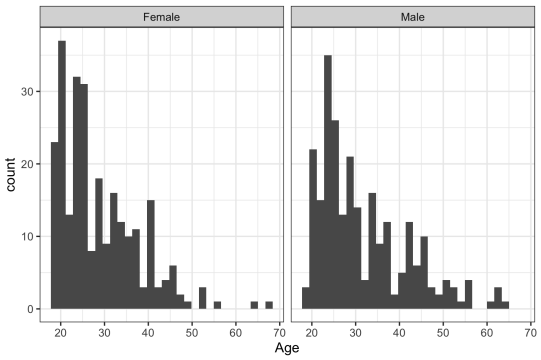

Figure 3: Histogram of Age counts by Gender.

2.2 Correlation

Correlation is a measure of association between two variables. We can obtain a correlation matrix between all variables using the `cor()` command in R. Each value in the matrix indicates the correlation between two given variables, with a value of 1 indicating a strong positive correlation (as A increases so does B), -1 indicating a strong negative correlation (as A increases, B decreases) and 0 indicating no correlation. Figure 4 displays positive correlations between a majority of variables i.e. an increase in the value of one variable leads to an increase in other. We identify a correlation of -0.08 for Thigh.girth and Gender, and a value of -0.02 for Thigh.girth and Age. These values would suggest no correlation between these variables.

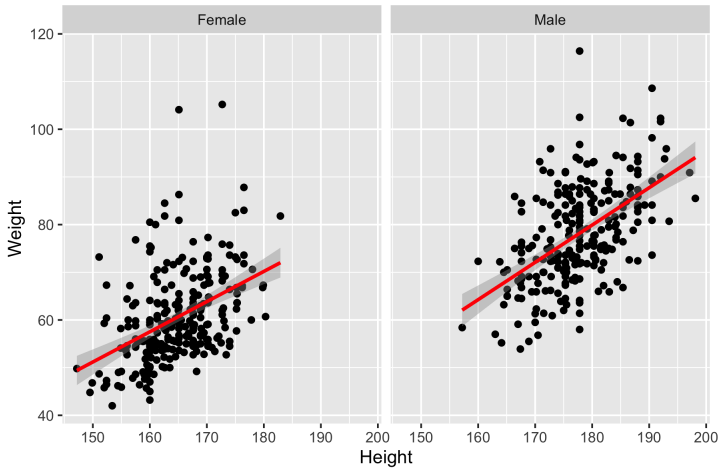

Figure 5: Scatter plot of weight against height per gender.

The plot of weight against height in Figure 5 agrees with our intuition, that is, people who are taller tend to be heavier due to an increase in volume with height. For both genders we generally notice a similar positive correlation of weight and height, although we notice a slightly steeper gradient for males compared to females. We can see the effect that height has on the weight of an individual differs according to the gender of the individual. This could indicate an interaction effect and we will explore this further in the modelling section.

2.3 Descriptive Statistics

Basic summary statistics for each variable, generated using the `stargazer` package in R, are listed in Appendix B. From these results we can see all values of `N` are 507, indicating we have 507 observations. Our response variable, Weight, has a mean of 69.148kg and a standard deviation of 13.346kg. We can calculate our range by subtracting the minimum value (42kg) from the maximum value (116.4kg) to obtain a range of 74.4kg. We can identify Shoulder, Chest, Waist and Weight

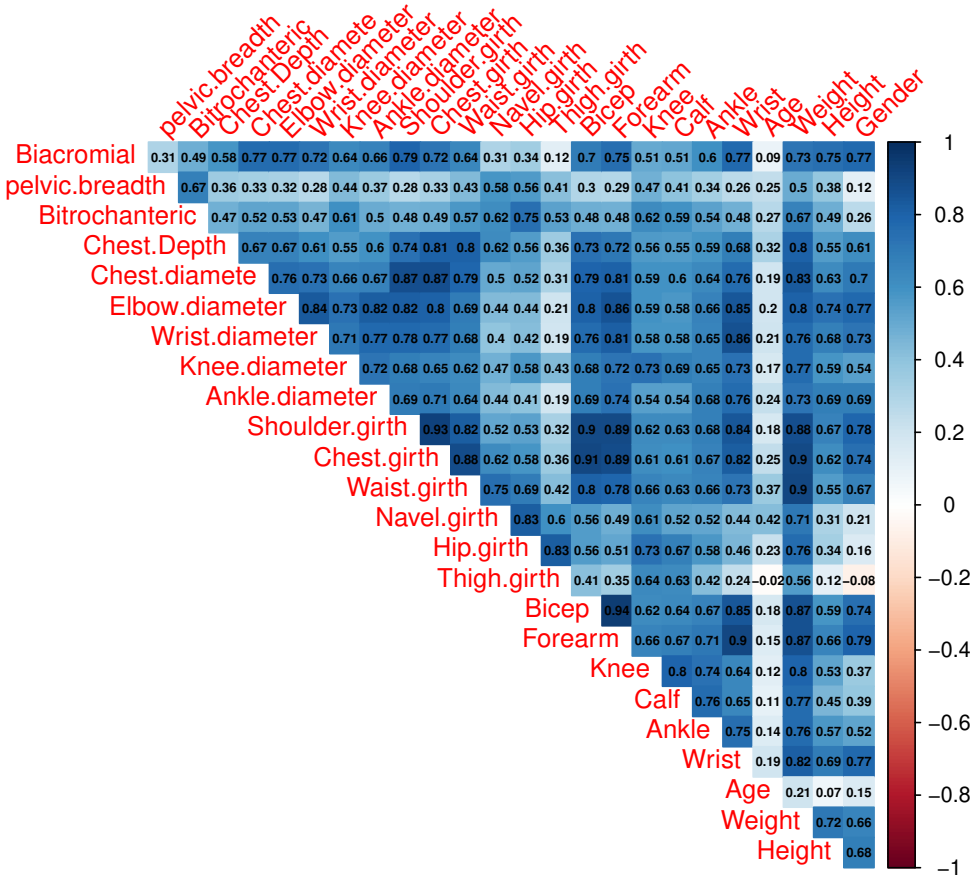

Figure 4: Correlation between all continuous variables

variables have large standard deviations, indicating the measurements recorded for these variables are widely spread out from the mean value compared to the other variables.

2.4 Summary

From our exploratory findings we can conclude almost all variables are correlated with each other with only a few exceptions of no correlation. This intuitively makes sense, as a large skeletal diameter measurement will result in other skeletal measurements also being larger as a consequence. This meant many of our exploratory plots resulted in scatter plots with positive correlations. We also found that transforming continuous variables to aid interpretation was not required.

3 Modelling

3.1 Main effects model

A sequence of Normal linear regression models were fitted to the data, treating Weight as the response variable and the remaining variables as explanatory variables.

Formally, if y_i is the weight of participant i, then the model fitted is:

$$y_i = \beta_0 + \sum_j \beta_j x_{ij} + \epsilon_i$$

where x_{ij} is the value of the jth covariate for person i, with $\epsilon_i \sim N(0, \sigma^2)$.

In order to find the most parsimonious model (the one which balances minimal assumptions with the greatest explanatory power), we begin with the full main effects model (`lm(Weight~.)`)

and carry out model selection using backwards elimination. This was implemented via R's `step()` function, which repeatedly identifies the least useful term based on the Akaike information criterion (AIC), until we are left with a model where all terms are significant.

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -122.54162    2.80227 -43.729  < 2e-16 ***
pelvic.breadth    0.11008    0.06217   1.770 0.077435 .
Chest.Depth       0.34331    0.08042   4.269 2.48e-05 ***
Chest.diamete     0.26028    0.08703   2.991 0.002960 **
Elbow.diameter    0.30564    0.19076   1.602 0.109929
Knee.diameter     0.41534    0.14033   2.960 0.003268 **
Ankle.diameter    0.28929    0.16109   1.796 0.073310 .
Shoulder.girth    0.09147    0.03336   2.742 0.006385 **
Chest.girth       0.08166    0.04091   1.996 0.046661 *
Waist.girth       0.38930    0.02805  13.878  < 2e-16 ***
Hip.girth         0.18732    0.04303   4.353 1.72e-05 ***
Thigh.girth       0.29623    0.05526   5.360 1.43e-07 ***
Bicep             0.16066    0.09168   1.752 0.080515 .
Forearm           0.30076    0.14141   2.127 0.034058 *
Knee              0.14868    0.08520   1.745 0.081756 .
Calf              0.25877    0.06686   3.870 0.000128 ***
Age              -0.05871    0.01348  -4.354 1.71e-05 ***
Height            0.30136    0.01929  15.619  < 2e-16 ***
GenderMale       -1.89928    0.53304  -3.563 0.000412 ***

Residual standard error: 2.034 on 386 degrees of freedom
Multiple R-squared:  0.979,Adjusted R-squared:  0.9781
F-statistic:  1002 on 18 and 386 DF,  p-value: < 2.2e-16
```

Figure 6: Output of the backwards elimination on the main effects model, after stepwise selection. L-R: variable name, regression coefficient, standard error, t-value, p-value

3.1.1 Regression Coefficients

Figure 6 displays the output of the main effects model after using the `summary()` command. The first row represents the intercept - the expected mean weight if all variables were to equal zero. Where there is a categorical variable, we take one possible value and consider that the baseline - here we assume gender is female. Taking into account the context of the data, we can say that a situation where all variables equal zero is never going to arise naturally, so the value of this the intercept coefficient has no useful meaning on its own.

However, the remaining coefficients tell us the additive effect that the other variables have on the response variable, using the intercept as a baseline value. For example, an increase in the pelvic breadth by one unit is associated with an increase of 0.095 in the weight, provided that all other explanatory variables stay constant. This confirms our findings from the correlation analysis that an an increase in the size of any bone is associated with an increase in weight. However we can now quantify the effect of each individual bone on weight - for example we can now say that chest depth has the most significant additive effect on weight.

Age displays a negative coefficient. This is logical since all of the participants were adults, and we would expect adults at the upper end of the age range to begin to weigh slightly less. The negative effect is not more significant as a result of the sample - the vast majority of participants were aged 20-30, with only a few at the upper end.

A more unexpected value is the negative coefficient for male participants, since we would generally expect men to weigh more than women. As discussed earlier however, there were many flaws in the sampling process and we therefore cannot expect the dataset's values to always reflect what we would expect of a wider population.

3.1.2 Standard Errors and the R^2 Statistic

Standard error and R^2 are statistics used to measure goodness of fit, or the extent to which the model fits the specific values in this dataset. Most of the standard error values are relatively small compared to their corresponding coefficients. Gender displays a slightly higher standard error, but this is to be expected since the variable only has two possible values, meaning that there will be a greater variation of weights within each gender category.

R^2, or the *coefficient of determination* measures the proportion of the variance explained by the model, by comparing the current fitted model with a model with just the mean fitted. For our model, $R^2 = 0.979$, meaning that 97.9% of variation is explained and therefore indicating that the model is an excellent fit [4]. The `summary()` command also produces an adjusted R^2 statistic, a modified version of R^2 adjusted for the number of predictors in the model, which aims to eliminate the risk of the statistic spuriously increasing with the number of variables. The adjusted R^2 indicates the percentage of variation explained by only the independent variables that actually affect the response variable. However, the adjust R^2 value for this model was 0.978, indicating that our model remains a good fit.

3.1.3 *p*-values

The p-value represents the probability of the given coefficient occurring under the null hypothesis, that the variable does not affect weight and the coefficient value is zero ($\beta = 0$). We usually test at the 5% level, meaning that a p-value < 0.05 is considered sufficient to reject the null hypothesis. We obtain our p-values by first calculating a t-value, which is computed by dividing the coefficient by the standard error. The p-value is then derived from a t-distribution and checked against our significance level of 0.05, and is considered statistically significant if this p value is smaller than our significance level. From the results we notice that the majority of variables kept after backwards elimination are statistically significant, that is, there is enough evidence to suggest they contribute to the weight of an individual. Those variables which may not seem statistically significant according to their p-value are left in the model, since they have been judged to be useful according to our model selection criterion, the AIC.

3.1.4 Residual deviance

When fitting a linear regression model we aim to minimise the residual sum of squares (also known as *deviance*). This value shows how well the response variable is predicted by the model when the predictor variables are included. For our model, R's `deviance()` function return a value of 1597.35.

3.2 Interactions

An interaction effect occurs when the value of one particular variable is dependent on the value of another. We utilise forward selection to discover interactions. Forward selection is a stepwise regression technique which starts with the simplest possible model (a model which only fits a mean to all observations) and iteratively tests potential interactions and adding those that contribute to the model's predictive power. This approach is far more appropriate for our dataset than the alternative of backwards selection, since the large number of variables would make computing and testing every possible interaction effect time-consuming and computationally expensive. In this section we report the results of our forward selection and discuss two particular interactions.

```
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                24.625735  20.424194   1.206 0.228705
    ...                                                    ...
Chest.Depth:Hip.girth       0.031186   0.008901   3.504 0.000515 ***
Knee.diameter:Shoulder.girth  0.028111   0.010953   2.567 0.010667 *
Calf:Age                   -0.012854   0.005018  -2.562 0.010809 *
Thigh.girth:GenderMale     -0.022139   0.077938  -0.284 0.776525
Chest.girth:Waist.girth     0.008931   0.002079   4.295 2.24e-05 ***
Height:GenderMale           0.074238   0.033387   2.224 0.026788 *
Chest.diamete:Bicep        -0.035477   0.015164  -2.339 0.019847 *
Ankle.diameter:Shoulder.girth  0.039534   0.014420   2.742 0.006413 **
Knee.diameter:Ankle.diameter  -0.114545   0.077478  -1.478 0.140149
Waist.girth:GenderMale     -0.069655   0.053070  -1.312 0.190172
Age:Height                  0.003146   0.001577   1.995 0.046735 *
Waist.girth:Age            -0.002615   0.001753  -1.492 0.136525
Chest.diamete:Age           0.022978   0.008169   2.813 0.005176 **
Chest.diamete:Hip.girth    -0.015744   0.009128  -1.725 0.085413 .
Knee.diameter:Age          -0.020997   0.010720  -1.959 0.050907 .
pelvic.breadth:Knee         0.051972   0.020405   2.547 0.011272 *
Knee:GenderMale            -0.251199   0.139375  -1.802 0.072315 .
pelvic.breadth:Chest.diamete  -0.043198   0.022793  -1.895 0.058849 .
Chest.girth:Age            -0.004055   0.002736  -1.482 0.139269

Residual standard error: 1.736 on 367 degrees of freedom
Multiple R-squared:  0.9855,Adjusted R-squared:  0.984
F-statistic: 673.5 on 37 and 367 DF,  p-value: < 2.2e-16
```

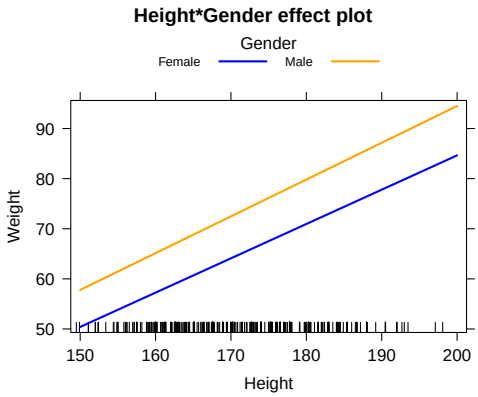

Figure 7: Effect plot of the interaction `Weight~Height:Gender`

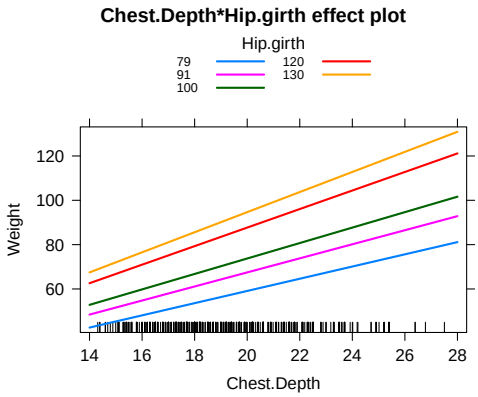

Figure 8: Effect plot of the interaction `Weight~Chest.Depth:Hip.girth`

In our exploratory data analysis, we identified height and gender as the source of a potential interaction effect. Our stepwise analysis confirms that there is a very slight interaction here (coefficient = 0.074, $p = 0.027$). Figure 8 displays an example of a more dramatic interaction, between chest depth and hip girth. There is a visible, albeit still minor, difference in gradient in different categories of hip girth.

It has been observed that spurious interaction terms can occur as a result of high levels of multicollinearity [5]. This is potentially relevant to our dataset, since we might intuitively expect all of the variables to be dependent on each other and positively correlated to some degree as a result of human growth. Our exploratory analysis also established some form of positive correlation between the majority of the variables. In future studies it might be possible to identify further unseen relationships by computing higher order interactions, but for now we do not consider these interactions to add any useful information to our model.

4 Diagnostics

Fitting a model requires us to make as number of assumptions about our data, which may not all be true. The purpose of regression diagnostics is to verify whether these assumptions are true, and allow us to correct any discrepancies. Many diagnostic methods make use of *residuals* - the difference between an observed value and the value as predicted by the model. In this section, we discuss the linearity, normality, heteroscedasticity and presence of outliers in our dataset, as well as examine the ability of our model to make predictions from unseen data.

4.1 Linearity

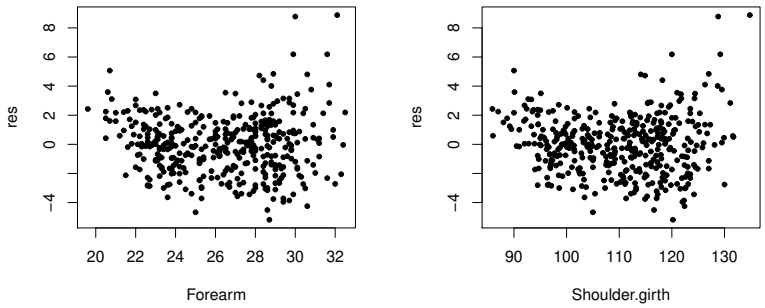

Figure 9: Two explanatory variables which display some non-linearity when plotted against the residuals

We can check the effectiveness of our model formula by plotting residuals against explanatory variables. Visible curvature would suggests that extra squared terms might need to be added. Two of the variables appeared to display some very slight non-linearity (Figure 9). In order to

test whether adding terms to our model might improve the fit, we fit two models: Model A, a main effects model of all variables; and Model B, the same model but with `Shoulder.girth` and `Forearm` replaced with `Shoulder.girth^2` and `Forearm^2` respectively. Model A had a residual standard error of 2.045, and an adjusted R^2 value of 0.977, whilst Model B had a standard error of 1.93 and adjusted R^2 of 0.982. We can conclude that these two models are almost identical in their fit, and any non-linearity in these two variables is not significant enough to influence our model.

4.2 Normality

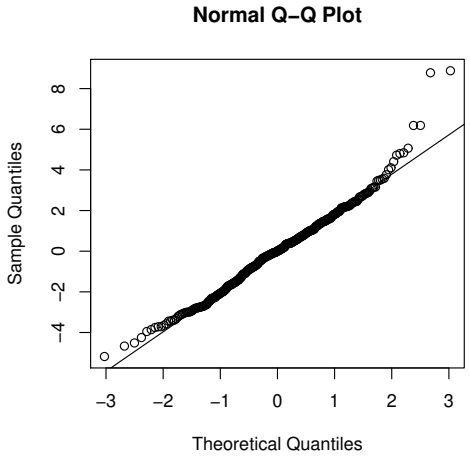

Figure 10: Quantile-quantile plot of the model's residuals against normally distributed residuals

We can verify that the residuals are normally distributed by plotting them against a set of residuals known to be normally distributed. A straight line indicates perfect normality, whilst large residuals a long way from the line would represent outliers. Figure 10 shows that the majority of the residuals fit the normal distribution, with no major outliers, demonstrating that the residuals are normally distributed. We can therefore conclude that our values are homoscedastic (i.e. have constant variance).

4.3 Homogeneity of Variance

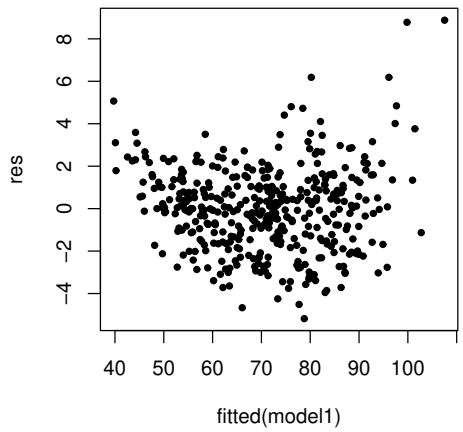

Figure 11: Plot of residuals against fitted values

When performing regression, we make the assumption that our data is *homoscedastic*, i.e. the variance is the same throughout the dataset. We can confirm this assumption visually by plotting the residuals against the fitted values (Fig. 11). A "fanning out" of the points from left to right or right to left would indicate heterogeneity of variance, but this is not the case.

4.4 Outliers

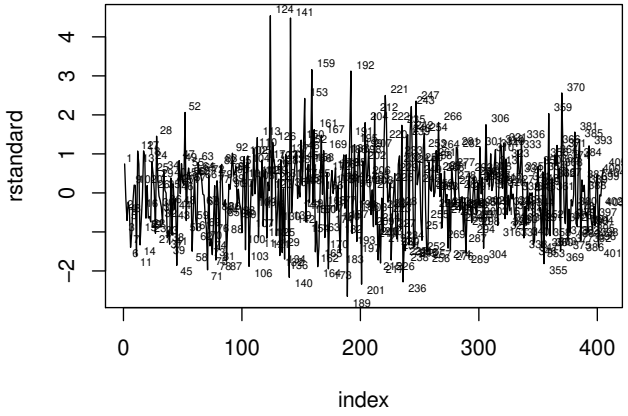

Figure 12: Plot of standardised residuals against index

We can find outliers by looking for points with unusually large residuals. Figure 12 shows the standardised residuals, plotted against their index. We would expect around one in twenty residuals will be outside of the range $[2, -2]$, so the number of outliers here does not seem unusual. Closer inspection of the most dramatic outliers (indices 124 and 141) reveals no massively irregular values that might indicate measurement error. If these measurements are genuine, it is likely that a model will run into more such values when making predictions on unseen data, and we therefore leave this very small number of outliers in the dataset to avoid overfitting our model. None of the other diagnostic plots discussed in this section would offer different conclusions if one to three points were to move, so we can say that these points have not affected the overall outcome of our analysis.

4.5 Transformation

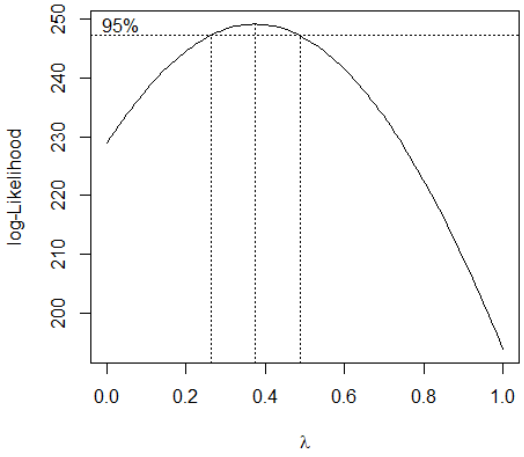

Figure 13: Box-cox estimate of λ

In order to determine whether our response variable requires transformation, we can use the Box-Cox family of transformations:

$$T(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ log(y) & \text{if } \lambda = 0 \end{cases}$$

By testing a range of λ values, we can use maximum likelihood to find the optimum transformation, i.e. the one which produces a distribution most resembling a normal distribution [4]. The optimum λ found was approximately 0.4 (Fig. 13), which corresponds to transforming the Weight variable by its 4th root. Whilst the adjusted R^2 remained unchanged, the new model with a transformed

response variable gave a residual standard error of 0.019, compared to 2.05 for the original model, indicating that this transformation has improved the fit of the model.

4.6 Prediction accuracy

Model testing can be either *in sample* or *out of sample*. Out of sample validation, also known as *cross-validation* refers to testing the new model against a set of unseen data, i.e. data that the model was not fitted on. This is in contrast to in sample testing, where we use data that the model has already "seen". The out of sample approach is generally considered to be more useful, and a better test of prediction quality, since it gives the best indication of how the model would perform in a practical scenario. We therefore split our dataset into two: the training set containing 80% of the original data (405 entries), and the testing set containing 20% (102). The model fitting described in the previous sections was carried out with this training set. When the initial model produced in Section 3 was used to make predictions from the test set, the mean error (distance between actual and predicted weight) was 1.96kg. We then applied the improved model, based on the diagnostics, and found the mean error to be 1.78, indicating a marginal improvement in the model's predictive ability.

5 Conclusions

In this paper we successfully implemented a linear model to predict an individual's weight based on multiple explanatory variables, and were successful in our original aim of determining the key skeletal and girth measurements which contribute to a person's weight, as well quantifying the size of these effects. This was achieved in the modelling section, where we used stepwise model selection on the dataset to achieve the most parsimonious model. We also discovered some interactions, with a slight variation in the weight of an individual based on their height differing by gender, and an example of the chest depth of an individual affecting weight differently based on the hip girth measurement, although the effects of both of these were very minor. We carried out regression diagnostics on our fitted model to check the assumptions made in the modelling section all hold and found no assumptions were violated. We then applied a Box-Cox transformation to our response variable and found an improvement in the model fit.

We envisage future work could involve more extensive prediction and testing. Computing higher order interactions would help to identify further unseen relationships. Implementing the K-Folds cross validation technique would ensure every observation from the original dataset has the chance of appearing in the training and testing sets.

References

[1] G. Heinz, L. J. Peterson, R. W. Johnson, and C. J. Kerk, "Exploring relationships in body dimensions," *Journal of Statistics Education*, vol. 11, no. 2, 2003.

[2] A. R. Behnke and J. H. Wilmore, *Evaluation and Regulation of Body Build and Composition.* Prentice Hall, 1974.

[3] NASA Scientific and Technical Information Office, *Anthropometric Source Book Volume 2 A Handbook of Anthropometric Data.* National Aeronautics and Space Administration, 1978.

[4] A. J. Dobson, *An Introduction to Generalized Linear Models*, 2nd ed. Chapman & Hall / CRC Press, 2002.

[5] J. M. Cortina, "Interaction, nonlinearity, and multicollinearity: Implications for multiple regression," *Journal of Management*, vol. 19, no. 4, pp. 915 – 922, 1993. [Online]. Available: http://www.sciencedirect.com/science/article/pii/014920639390035L

A Code

(code in electronic submission)

B Summary statistics

Table 2: Descriptive Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Biacromial	507	38.811	3.059	32.400	36.200	41.150	47.400
pelvic.breadth	507	27.830	2.206	19	26.5	29.2	35
Bitrochanteric	507	31.980	2.031	24.700	30.600	33.350	38.000
Chest.Depth	507	19.226	2.516	14.300	17.300	20.900	27.500
Chest.diamete	507	27.974	2.742	22	25.6	29.9	36
Elbow.diameter	507	13.385	1.353	9.900	12.400	14.400	16.700
Wrist.diameter	507	10.543	0.944	8.100	9.800	11.200	13.300
Knee.diameter	507	18.811	1.348	15.700	17.900	19.600	24.300
Ankle.diameter	507	13.863	1.247	9.900	13.000	14.800	17.200
Shoulder.girth	507	108.195	10.375	85.900	99.450	116.550	134.800
Chest.girth	507	93.334	10.028	72.600	85.300	101.150	118.700
Waist.girth	507	76.979	11.013	57.900	68.000	84.500	113.200
Navel.girth	507	85.654	9.424	64.000	78.850	91.600	121.100
Hip.girth	507	96.681	6.681	78.800	92.000	101.000	128.300
Thigh.girth	507	56.856	4.460	46.300	53.700	59.500	75.700
Bicep	507	31.170	4.247	22.400	27.600	34.450	42.400
Forearm	507	25.943	2.831	20	23.6	28.4	32
Knee	507	36.203	2.618	29.000	34.400	37.950	49.000
Calf	507	36.078	2.848	28.400	34.100	38.000	47.700
Ankle	507	22.157	1.862	16.400	21.000	23.300	29.300
Wrist	507	16.097	1.381	13.000	15.000	17.100	19.600
Age	507	30.181	9.608	18	23	36	67
Weight	507	69.148	13.346	42.000	58.400	78.850	116.400
Height	507	171.144	9.407	147	163.8	177.8	198