Statistics 1 Project

-ARYAN K, SAMITINJAYA P, MOHAIDEEN A.K, PIYUSH B

Introduction

The paper we are referring to is titled “Exploring Relationships in Body Dimensions”. It is authored by Grete Heinz, Louis Peterson, Roger Johnson, and Carter Kerk. It was published in the Journal of Statistics Education (Volume 11, Number 2). The paper has taken various measurements- body girth measurements, skeletal diameter measurements, age, height, weight, and gender. These measurements were taken for 507 physically active individuals- 247 men and 260 women. In our project, we will be performing analysis on this data.

Journal of Statistics Education

Exploring Relationships in Body Dimensions

Grete Heinz

Louis J. Peterson
San José State University

Roger W. Johnson
South Dakota School of Mines and Technology

Carter J. Kerk
South Dakota School of Mines and Technology

Journal of Statistics Education Volume 11, Number 2 (2003), jse.amstat.org/v11n2/datasets.heinz.html

---

Key Words: Anthropometry; Discriminant analysis; Ergonomics; Forensic science; Multiple regression.

Some Preliminaries

What The Paper Achieves

The paper's aim was to investigate the relationship between body build (leanness/fatness), weight, and girths in a group of physically active young men and women. Most of the men and women were within the normal weight range. Body build was determined using skeletal width and depth measurements taken at nine well-defined body sites. For the given sample, the paper was able to affirm the hypothesis that body build variables (skeletal variables) and height predict scale weight (measured weight) substantially better than height alone.

Following this, a weight equation was found for the group using linear regression. This equation gives the weight in terms of body build variables and height. So now, for each person, there are two weights to keep track of:

Scale Weight- This is the weight of the person as given by the scale.

Body Build Weight- This is the projected weight of the person, as per the group's weight equation.

The regression equation is

$$\begin{aligned}\text{Weight (kg)} = {} & -120 + 0.0781\ \text{Shoulder Girth} + 0.198\ \text{Chest Girth} \\ & + 0.340\ \text{Waist Girth} + 0.0012\ \text{Navel Girth} \\ & + 0.240\ \text{Hip Girth} + 0.314\ \text{Thigh Girth} + 0.0547\ \text{Flexed Bicep Girth} \\ & + 0.532\ \text{Forearm Girth} + 0.301\ \text{Knee Girth} + 0.404\ \text{Calf Maximum Girth} \\ & - 0.0096\ \text{Ankle Minimum Girth} - 0.118\ \text{Wrist Minimum Girth} \\ & + 0.328\ \text{Height}\end{aligned} \qquad (1)$$

Here, the parameters on the left side of the equation are all measured in centimetres.

Following this, trunk and limb girths were measured for each person from twelve well-defined body sites. Again, the authors used regression analysis to obtain best prediction equations for the measured girths from selected body build variables. Body build girths were then projected from these girth equations.

Data Sources

As mentioned, measurements were taken on 247 men and 260 women. These were primarily individuals in their twenties and early thirties, with a few older men and women. All these individuals were physically active (that is, they exercised for several hours per week). There is something we need to be cautious about, though. The dataset does not constitute a random sample from a well-defined population. So, there may be some inherent bias in the readings. For example, we are only considering physically active individuals. Moreover, we are not considering younger individuals. This is because younger individuals' skeletons are still developing and hence taking any measurements will not be consistent.

Description of the Data

Nine skeletal measurements (diameter measurements) were included. A broad-blade anthropometer was used to measure the biacromial, biiliac, bitrochanteric, and chest diameters along the trunk. A smaller anthropometer was used to for four skeletal measurements along the limbs- the elbow, wrist, knee, and ankle diameters. For measuring the chest depth, the depth attachment of the anthropometer was activated. Firm pressure was applied at each body site to effectively compress the flesh so that “bone-to-bone” measurements could be obtained. One thing to note is that at the time of maturation, all the skeletal sites have achieved their maximum size.

Twelve girth measurements are used in the study. In general, these measurements vary with time- with the exception of the three bony girths of the wrist, knee, and ankle. The changeable girths are the shoulder, chest, waist, navel, hip, thigh, bicep, forearm, and calf ones.

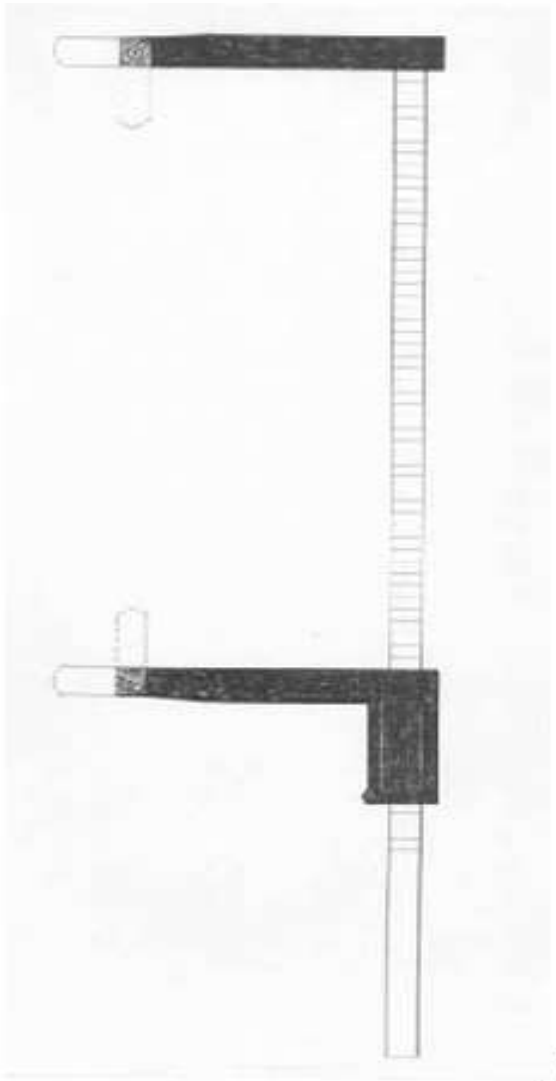


Figure 1

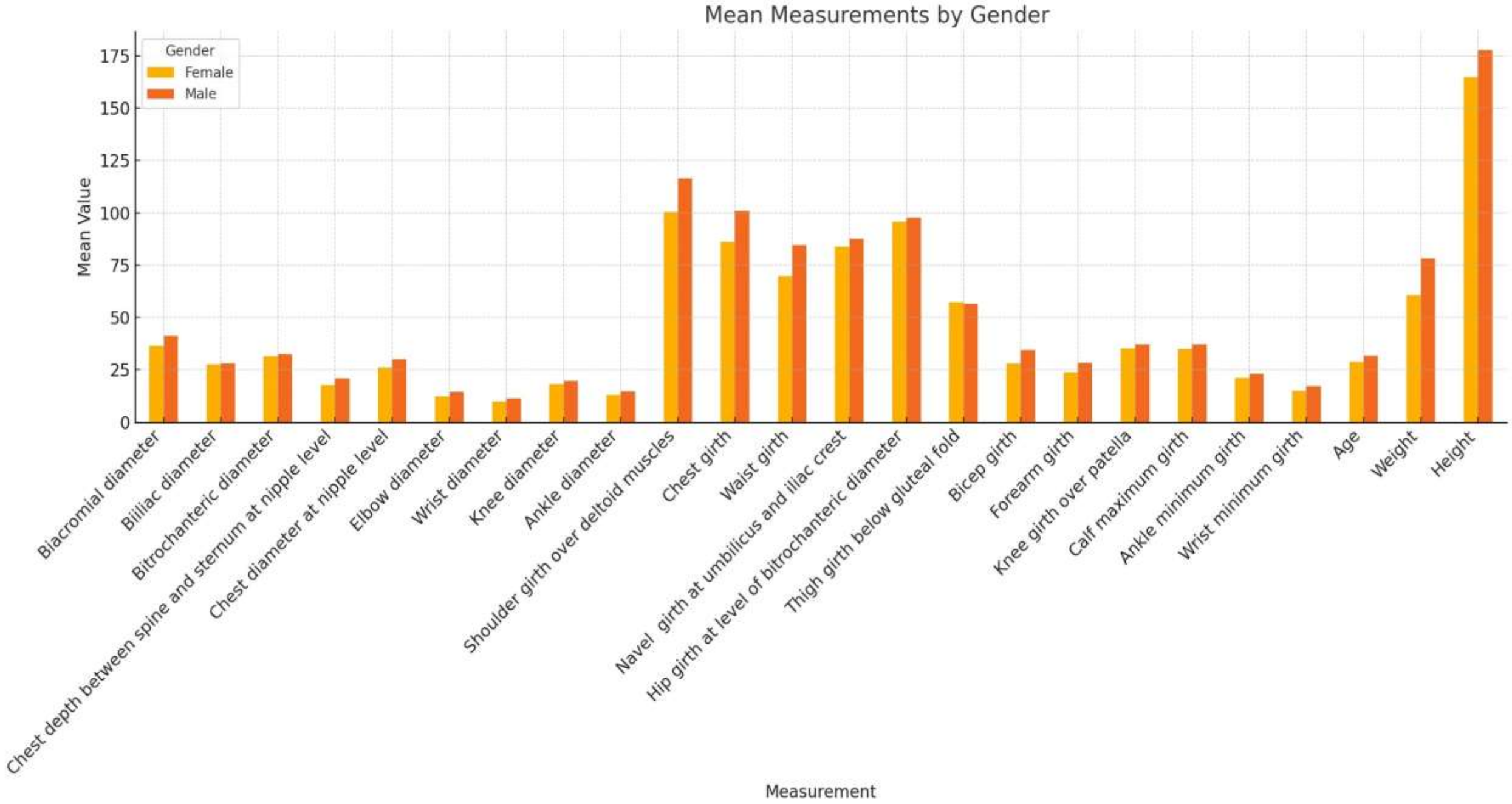
Figure 1. Sliding Anthropometer with Depth Attachment.

Anthropometer

An anthropometer is a specialized measuring device used to measure dimensions of the human body. It consists of a long, straight rod or series of rods with attached sliding calipers. This allows for accurate measurements of various body parts. The attachment at the end of the calipers help in depth measurement on parts which are otherwise surrounded by bigger skeletons or muscles on the outer sides.

Analysing the Data

	Men	Women	All
Biacromial diameter	: 41.2413	: 36.50308	: 38.81144
Biiliac diameter	: 28.0915	: 27.58154	: 27.82998
Bitrochanteric diameter	: 32.52672	: 31.46154	: 31.98047
Chest depth	: 20.80648	: 17.72462	: 19.22604
Chest diameter	: 29.94899	: 26.09731	: 27.97377
Elbow diameter	: 14.45709	: 12.36692	: 13.38521
Wrist diameter	: 11.24615	: 9.874231	: 10.5426
Knee diameter	: 19.56194	: 18.09692	: 18.81065
Ankle diameter	: 14.74413	: 13.02654	: 13.86331
Shoulder girth	: 116.5016	: 100.3038	: 108.1951
Chest girth	: 100.9899	: 86.06	: 93.33353
Waist girth	: 84.5332	: 69.80346	: 76.97949
Navel girth	: 87.66235	: 83.74577	: 85.65385
Hip girth	: 97.76316	: 95.65269	: 96.68087
Thigh girth	: 56.49798	: 57.19577	: 56.85582
Bicep girth	: 34.40364	: 28.09731	: 31.16963
Forearm girth	: 28.24049	: 23.76038	: 25.943
Knee girth	: 37.19555	: 35.26	: 36.20296
Calf girth	: 37.20688	: 35.00615	: 36.0783
Ankle girth	: 23.15911	: 21.20577	: 22.1574
Wrist girth	: 17.19028	: 15.05923	: 16.09744
Age	: 31.66802	: 28.76923	: 30.18146
Weight	: 78.14453	: 60.60038	: 69.14753
Height	: 177.7453	: 164.8723	: 171.1438

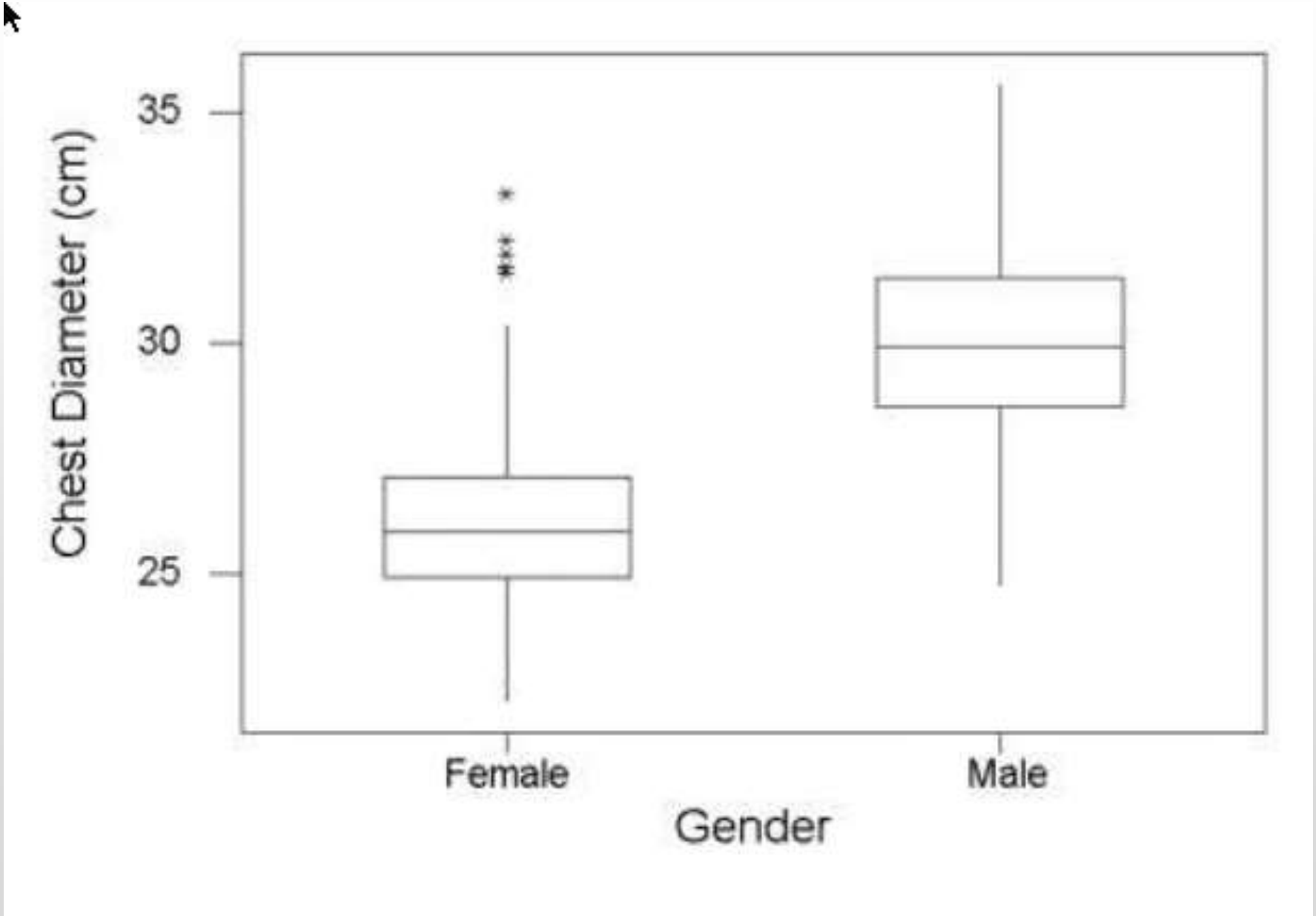

Mean Measurements by Gender
Gender
Female
Male
Mean Value
0
25
50
75
100
125
150
175
Biacromial diameter
Bilíiac diameter
Bitrochanteric diameter
Chest depth between spine and sternum at nipple level
Chest diameter at nipple level
Elbow diameter
Wrist diameter
Knee diameter
Ankle diameter
Shoulder girth over deltoid muscles
Chest girth
Waist girth
Navel girth at umbilicus and iliac crest
Hip girth at level of bitrochanteric diameter
Thigh girth below gluteal fold
Bicep girth
Forearm girth
Knee girth over patella
Calf maximum girth
Ankle minimum girth
Wrist minimum girth
Age
Weight
Height
Measurement

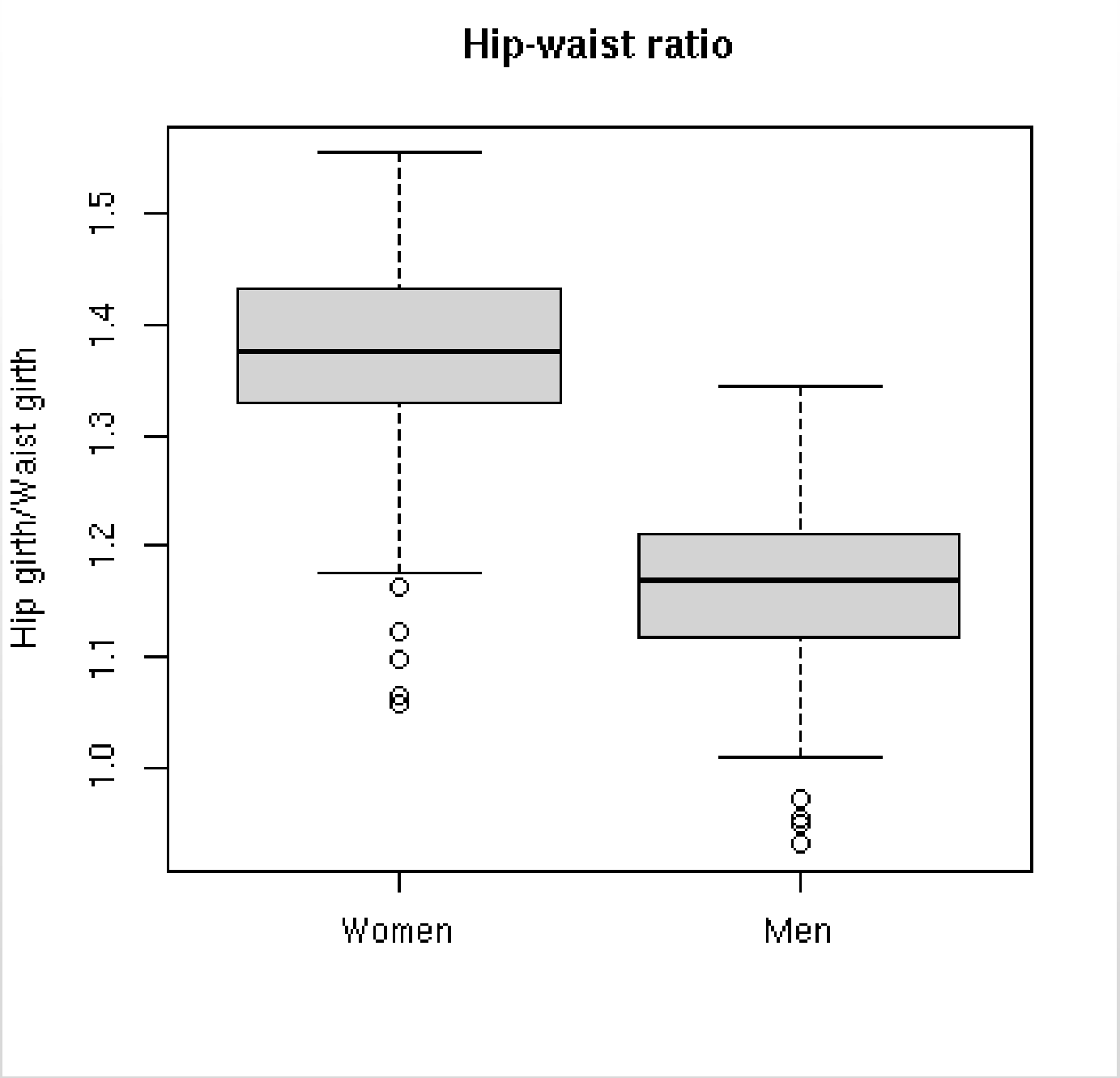
Inferences

Using the graph from the previous slide, we infer that usually men have larger measurements than women. In some measurements like the biiliac diameter, the gap is not so apparent. However, one thing to note is that there is one measurement in which women exceed men- the thigh girth.

Some Box Plots

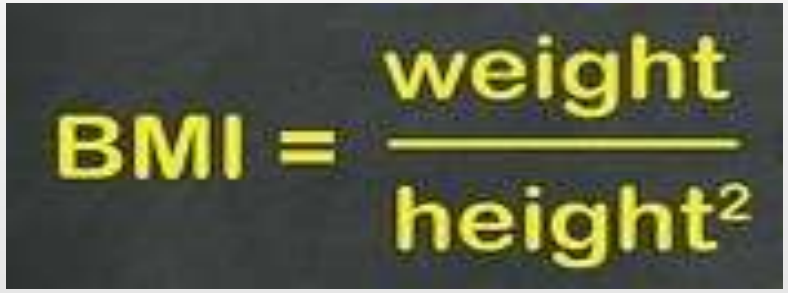

Comparison of bicep girths
Bicep girth(cm)
25
30
35
40
Men
Women


Chest Diameter (cm)
35
30
25
Female
Male
Gender


Hip-waist ratio
Hip girth/Waist girth
1.0
1.1
1.2
1.3
1.4
1.5
Women
Men

BMI

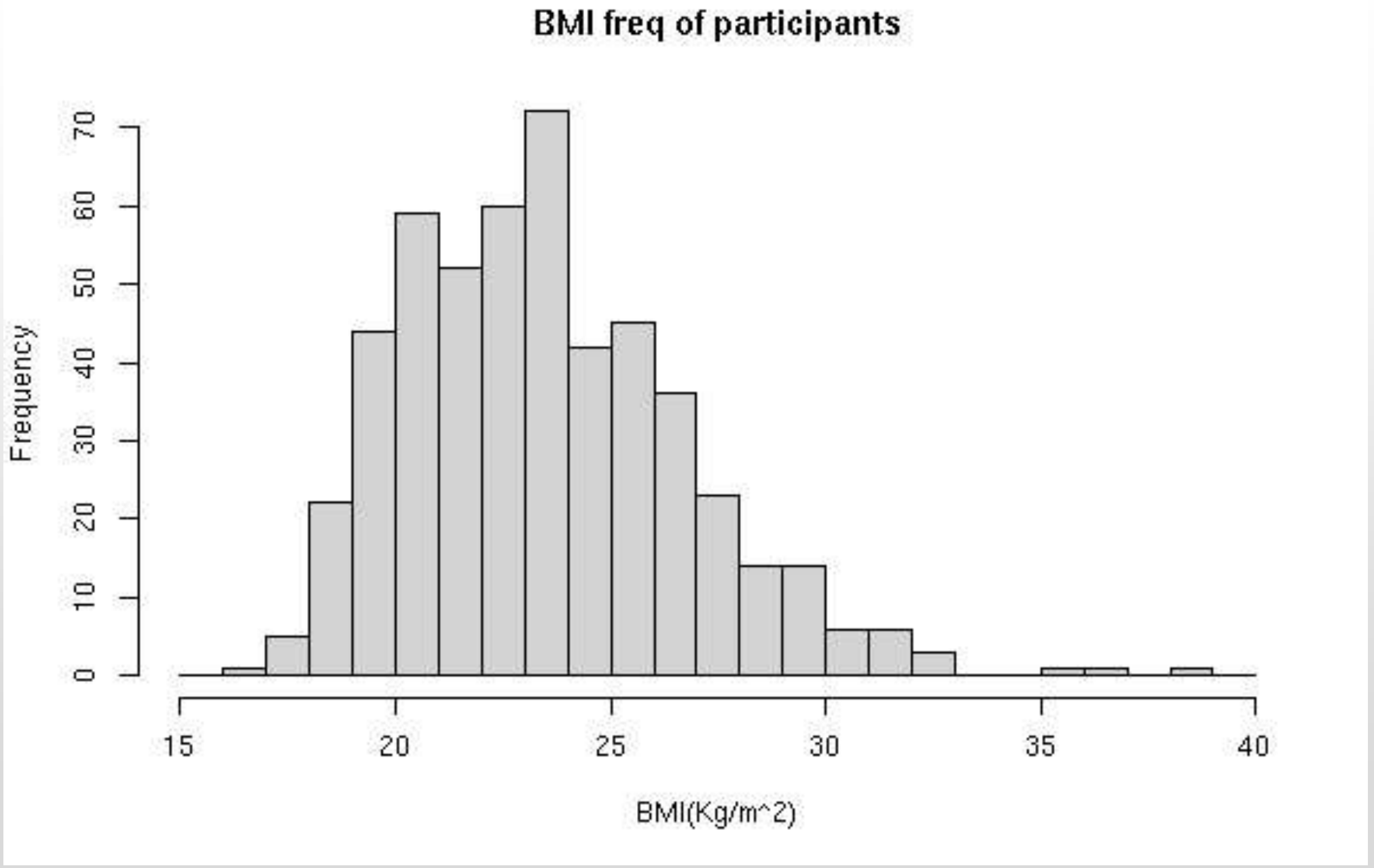
The Body Mass Index, or BMI is a measure of the body fat based on the height and the weight of an Adult. It is given by the formula:

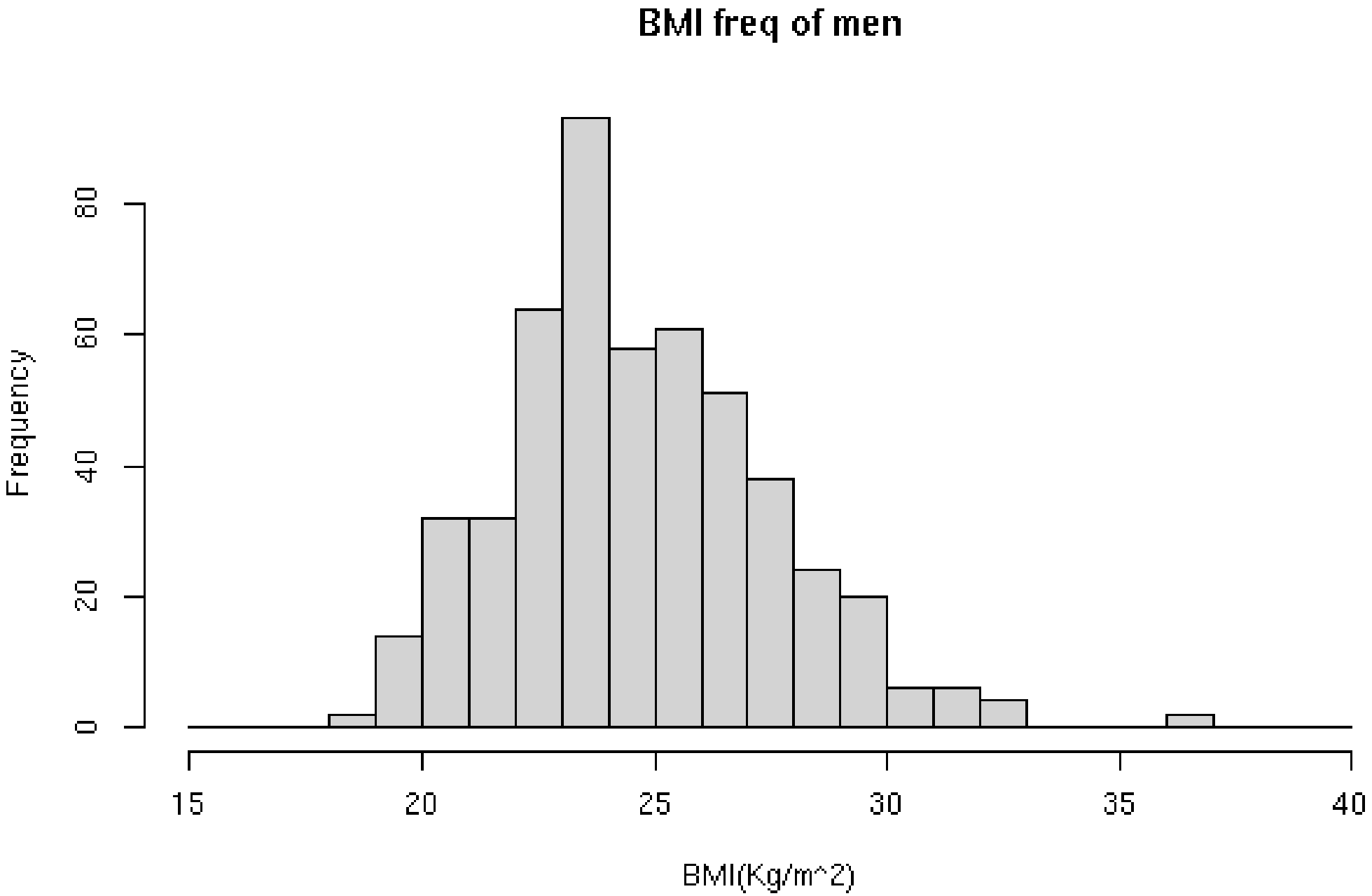


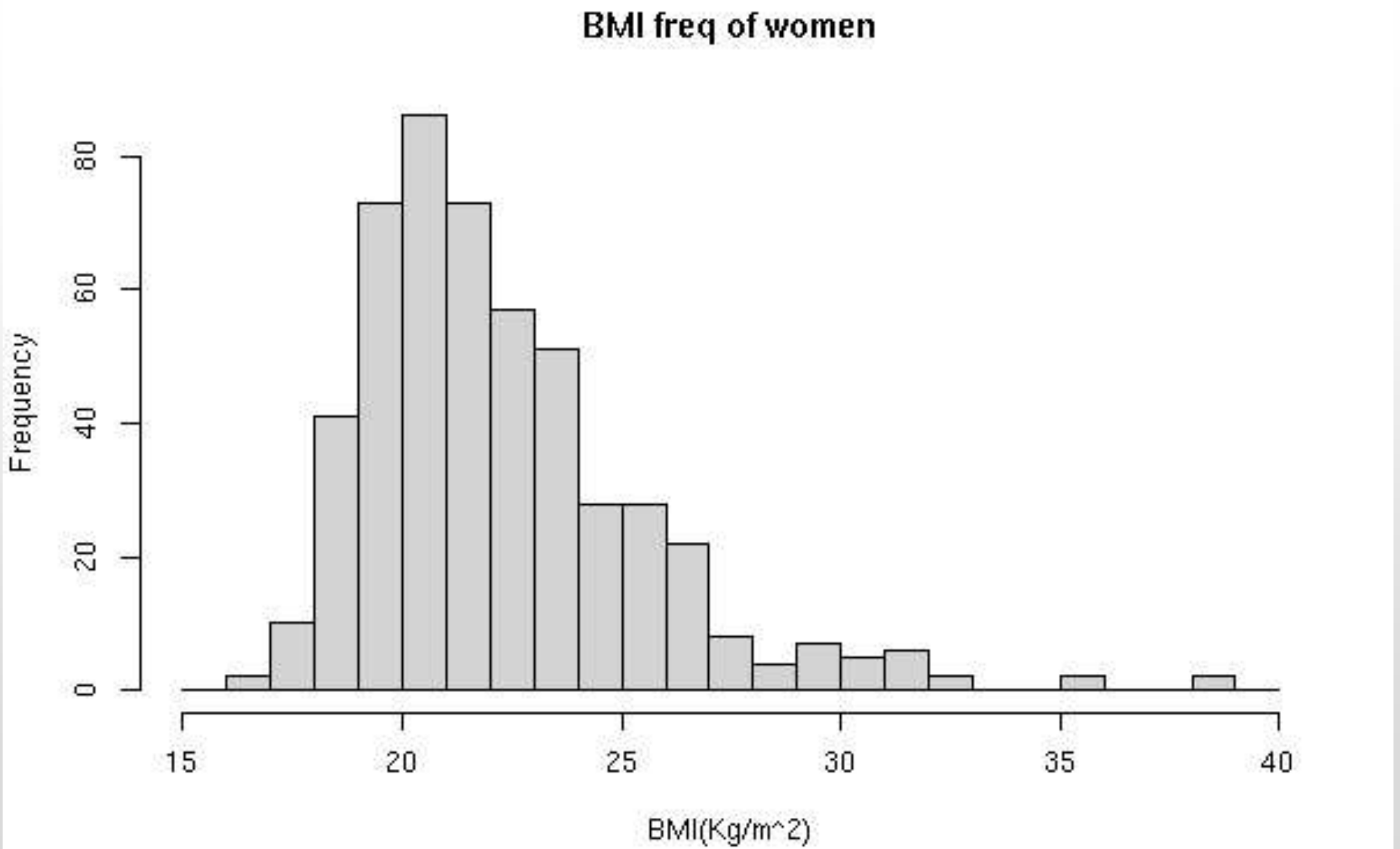
Where Weight is measured in Kg, and height is measured in metres.

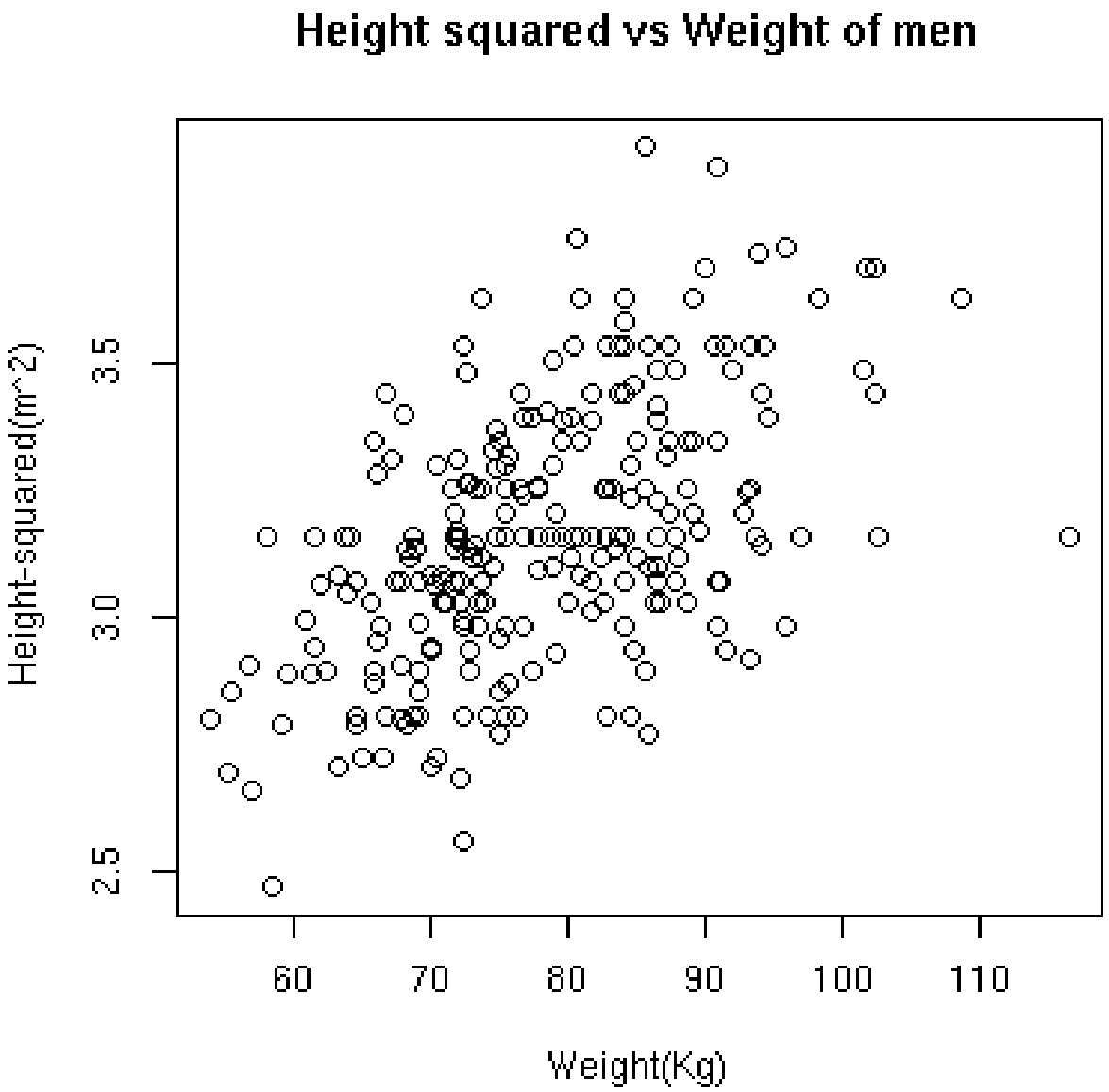
While not precise, it can give us a general evaluation of whether an individual has normal body fat levels, is underweight or obese.

We calculated the BMIs of the individuals and got the following results.


BMI freq of participants
Frequency
0
10
20
30
40
50
60
70
15
20
25
30
35
40
BMI(Kg/m^2)


BMI freq of men
Frequency
0
20
40
60
80
15
20
25
30
35
40
BMI(Kg/m^2)


BMI freq of women
Frequency
0
20
40
60
80
15
20
25
30
35
40
BMI(Kg/m^2)


Height squared vs Weight of men
Height-squared(m^2)
2.5
3.0
3.5
60
70
80
90
100
110
Weight(Kg)


Height squared vs Weight of women
Height-squared(m^2)
2.2
2.4
2.6
2.8
3.0
3.2
40
50
60
70
80
90
100
Weight(Kg)

The average BMI came up to be 23.84, with that of males only being 24.71 and the same for females being 22.28.

This shows that males tend to have a higher weight with respect to their height in comparison to females. This is due to the bigger skeletal and muscular measures they generally have as compared to their counterparts which we previously compared and saw.

CV

CV stands for the coefficient of variation. The formula is as follows:

$$\mathrm{CV} = \left(\frac{\sigma}{\mu}\right) \times 100$$

Here, sigma is the standard deviation of the body measurements and mu is the mean of the body measurements. The CV is unitless and it is a measure of the inherent variability of the body dimensions.

In the following section, we calculated the CVs of the measurements given to us and tabulated them.

CV	Men	Women	Overall
Biacromial diameter	: 5.06086	: 4.874167	: 7.882037
Biiliac diameter	: 7.358448	: 8.366016	: 7.92781
Bitrochanteric diameter	: 5.73415	: 6.513281	: 6.350486
Chest depth	: 10.30272	: 10.33627	: 13.08578
Chest diameter	: 6.95552	: 6.969332	: 9.800788
Elbow diameter	: 6.104572	: 6.762982	: 10.10747
Wrist diameter	: 5.654353	: 6.700543	: 8.957569
Knee diameter	: 5.476761	: 6.55692	: 7.164001
Ankle diameter	: 6.404147	: 6.648451	: 8.997496
Shoulder girth	: 5.577625	: 6.450997	: 9.589009
Chest girth	: 7.138357	: 7.169889	: 10.74386
Waist girth	: 10.3891	: 10.87016	: 14.306
Navel girth	: 9.564971	: 11.89509	: 11.00257
Hip girth	: 6.370542	: 7.256176	: 6.909974
Thigh girth	: 7.516495	: 8.105501	: 7.844209
Bicep girth	: 8.667796	: 9.643192	: 13.62525
Forearm girth	: 6.300617	: 7.080067	: 10.91076
Knee girth	: 6.110944	: 7.311631	: 7.230265
Calf girth	: 7.109264	: 7.464777	: 7.893001
Ankle girth	: 7.466123	: 6.785033	: 8.405036
Wrist girth	: 5.282035	: 5.640461	: 8.578576
Age	: 32.05584	: 30.77312	: 31.83568
Weight	: 13.45314	: 15.86739	: 19.30042
Height	: 4.041528	: 3.969497	: 5.496668

In the text “Bodyspace- Anthropometry, Ergonomics, and the Design of Work” written by Stephen Pheasant, there is a table of CVs for various body measurements (1996, Table A3, Page 219). We will now compare these CVs with the CVs we obtained from our dataset and see what conclusions we may draw.

Table A3 Characteristic coefficients of variation of anthropometric data.

Dimension	CV (%)
Stature	3–4
Body heights (sitting height, elbow height, etc.)	3–5
Parts of limbs	4–5
Body breadths (hips, shoulder, etc.)	5–9
Body depths (abdominal, chest, etc.)	6–9
Dynamic reach	4–11
Weight	10–21
Joint ranges	7–28
Muscular strength (static)	13–85

We see in the above table that body breadths have a standard CV of 5 to 9. In our dataset, the body breadths include Chest, Elbow, Wrist, Knee, and Ankle Diameters. We will now list the CVs for these following values:

Chest Diameter: 9.8

Elbow Diameter: 10.1

Wrist Diameter: 8.96

Knee Diameter: 7.16

Ankle Diameter: 8.99

In the table, stature (height) has a CV of 3 to 4. From our dataset, we get that height has a CV of 5.49.

In the table, weight has a CV of 10 to 21. From our dataset, we get that weight has a CV of 19.3.

We can see that the CVs from our datasets tend to exceed the “standard CVs” given in the table. This is to be expected. This is because our sample size is relatively small (507) and as discussed before, the sample is not representative of the general population. We also found from the dataset that the CV for the age is 31.83, which is exceptionally high. This could be an explanation for why our CVs exceed the “standard CVs”.

Recall from our earlier discussions that the authors of the paper had derived a regression equation for the weight:

The regression equation is

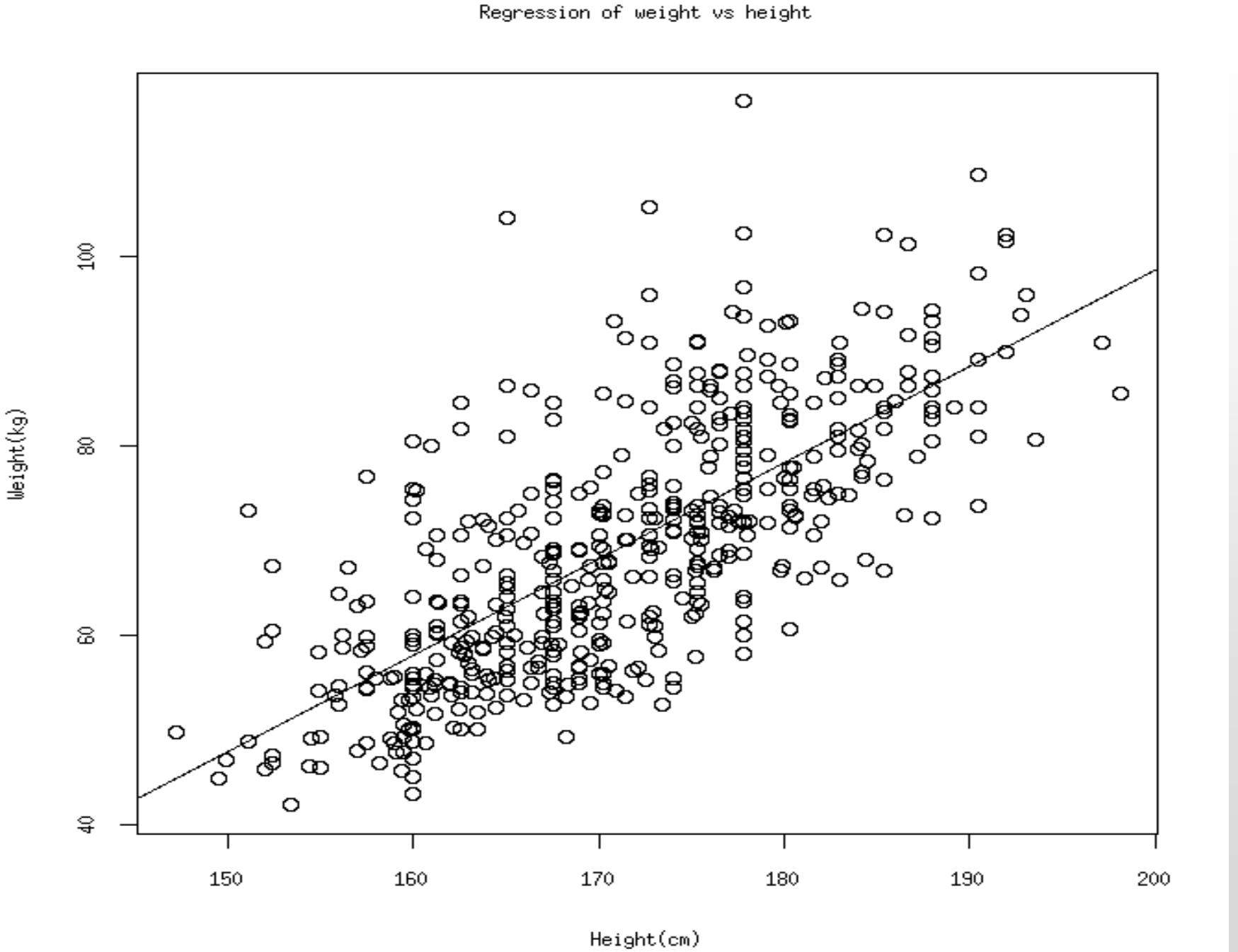
$$\begin{aligned}\text{Weight (kg)} = &- 120 + 0.0781\ \text{Shoulder Girth} + 0.198\ \text{Chest Girth} \\ &+ 0.340\ \text{Waist Girth} + 0.0012\ \text{Navel Girth} \\ &+ 0.240\ \text{Hip Girth} + 0.314\ \text{Thigh Girth} + 0.0547\ \text{Flexed Bicep Girth} \\ &+ 0.532\ \text{Forearm Girth} + 0.301\ \text{Knee Girth} + 0.404\ \text{Calf Maximum Girth} \\ &- 0.0096\ \text{Ankle Minimum Girth} - 0.118\ \text{Wrist Minimum Girth} \\ &+ 0.328\ \text{Height}\end{aligned} \qquad (1)$$

This equation gives a mean squared error of 4.72, a mean absolute error of 1.64 kg and a R^2 score of 0.973.

Now, to see how it stands, we did some linear regression on the heights and the weights of the individuals to get the following equation-

$$\text{weight} = (1.017617 \times \text{height}) - 105.0113$$

Here, weight is measured in kilograms and height is measured in centimetres.


Regression of weight vs height
Weight(kg)
40
60
80
100
150
160
170
180
190
200
Height(cm)

This equation gives a mean squared error of 86.30, a mean absolute error of 7.20 kg and a R^2 score of 0.515.

What is R^2 Score?

The R^2 score, also known as the coefficient of determination is the proportion of the variation in the dependent variable that is predictable from the independent variables. The closer the R^2 score is to 1, the better the model fits the data. The formula for R^2 is given in the next slide.

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

Here,

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$$

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$

Here, the y's are the measured values and the f's are the modelled values.

Clearly, the multivariable linear regression gives a much better model. The R^2 score is close to 1, and the MSEs and MAEs are also closer to 0. This verifies the hypothesis of the paper that the multivariable linear regression is a better estimator for the weight than the single variable linear regression model which only considers the height.

Our Analysis of the data

Another Multivariable Linear Regression

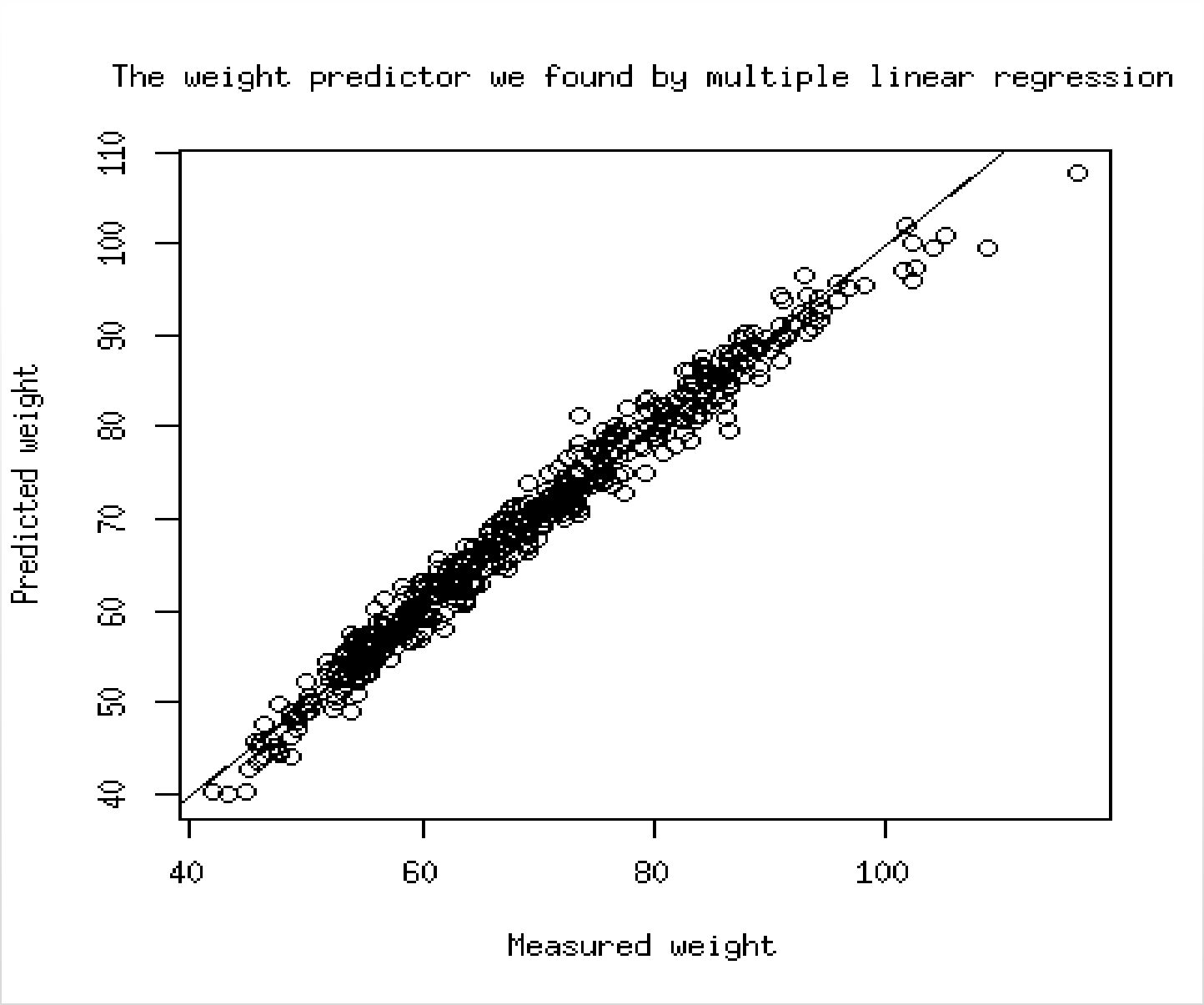
We decided to perform another multivariable linear regression. This time, we included age as an independent variable as well. The regression equation in this case is shown in the next slide.

```
Coefficients:
(Intercept)    bdat[, 1]    bdat[, 2]    bdat[, 3]    bdat[, 4]    bdat[, 5]
 -117.32855     -0.06456      0.12603     -0.07524      0.28114      0.16449
  bdat[, 6]    bdat[, 7]    bdat[, 8]    bdat[, 9]   bdat[, 10]   bdat[, 11]
    0.18488      0.21945      0.41815      0.04013      0.07258      0.13415
 bdat[, 12]   bdat[, 13]   bdat[, 14]   bdat[, 15]   bdat[, 16]   bdat[, 17]
    0.33766      0.01122      0.24837      0.26019      0.07770      0.40520
 bdat[, 18]   bdat[, 19]   bdat[, 20]   bdat[, 21]   bdat[, 22]   bdat[, 24]
    0.21695      0.36297      0.01190     -0.22194     -0.05542      0.28997
```

$$
\begin{aligned}
\text{Weight} = & -117.329 \\
& - 0.065 \cdot \text{Biacromial Diameter} \\
& + 0.126 \cdot \text{Biiliac Diameter} \\
& - 0.075 \cdot \text{Bitrochanteric Diameter} \\
& + 0.281 \cdot \text{Chest Depth} \\
& + 0.164 \cdot \text{Chest Diameter} \\
& + 0.185 \cdot \text{Elbow Diameter} \\
& + 0.219 \cdot \text{Wrist Diameter} \\
& + 0.418 \cdot \text{Knee Diameter} \\
& + 0.040 \cdot \text{Ankle Diameter} \\
& + 0.073 \cdot \text{Shoulder Girth} \\
& + 0.134 \cdot \text{Chest Girth} \\
& + 0.338 \cdot \text{Waist Girth} \\
& + 0.011 \cdot \text{Navel Girth} \\
& + 0.248 \cdot \text{Hip Girth} \\
& + 0.260 \cdot \text{Thigh Girth} \\
& + 0.078 \cdot \text{Bicep Girth} \\
& + 0.405 \cdot \text{Forearm Girth} \\
& + 0.217 \cdot \text{Knee Girth} \\
& + 0.363 \cdot \text{Calf Girth} \\
& + 0.012 \cdot \text{Ankle Girth} \\
& - 0.222 \cdot \text{Wrist Girth} \\
& - 0.055 \cdot \text{Age} \\
& + 0.290 \cdot \text{Height}
\end{aligned}
$$

This regression equation has a mean squared error of 4.25, a mean absolute of 1.59 kg and a R^2 score of 0.976. This is a slight improvement over the regression equation the authors provided. This model explains 97.6% of the variance in weight.

However, as our sample is not a depiction of the general population and there might be some inherent bias, this equation may be less accurate on other samples.


The weight predictor we found by multiple linear regression
Predicted weight
40
50
60
70
80
90
100
110
40
60
80
100
Measured weight

Some Variable Selection and Multiple Regression

Variable Selection

Variable Selection, also known as feature selection is the process of determining and selecting the important predictors for a model. This helps in getting a simpler and faster model in R and prevents overfitting. Overfitting occurs when the model is too complex and it makes it hard to capture the underlying pattern. It optimizes too much on the given sample which in turn makes it less efficient on other samples. Variable selection finds the most relevant variables, which helps improving the model's ability to generalize new data.

Techniques for Variable Selection in R:-

1) **Filter Methods**
2) **Wrapper Methods**
3) **Embedded Methods**

Filter Methods

Filter methods select important features without using a predictive model. They look at the data itself to decide which features matter most. Common methods include:

- **Chi-Square Test**: Used for categorical variables. It checks if two variables are related by comparing actual vs. expected values. Features with a p-value less than a chosen threshold are usually kept.
- **ANOVA (Analysis of Variance)**: Used when comparing group means. It checks if the difference between group averages is significant using an F-test. Features with a p-value below 0.05 are selected.
- **Correlation Coefficient**: Used for continuous variables. It measures how strongly two variables are related (using Pearson's correlation). Features with a high absolute correlation are chosen.

Wrapper Methods

Wrapper methods choose the best features by actually training and testing models with different feature combinations. They rely on a specific predictive model to see which features improve results the most. Common methods include:

- **Forward Selection**: Starts with no features, and adds them one at a time based on how much they help the model.
- **Backward Elimination**: Starts with all features, and removes them one by one if they don't help the model.
- **Stepwise Selection**: A mix of both—adds and removes features during the process, keeping only the ones that really improve performance.

These methods usually take more time because they involve training multiple models, but they can find better feature sets.

Embedded Methods

Embedded methods select features while building the model. The model itself decides which features are important during training. This makes them more efficient than wrapper methods. Common techniques include:

- **Lasso (L1 Regularization)**: Pushes less useful feature coefficients to zero, effectively removing them. Keeps features with non-zero values.
- **Ridge (L2 Regularization)**: Shrinks all coefficients but doesn't remove features. Features with smaller coefficients are less important.
- **Elastic Net**: Combines Lasso and Ridge. Useful when features are highly correlated.
- **Tree-Based Models** (like Decision Trees & Random Forests): Measure how useful a feature is by how well it splits the data. Features that cause big reductions in impurity (like Gini or entropy) are kept.

RFE

We used Recursive Feature Elimination, which is a wrapper method similar in functionality to Backward Elimination. It eliminates the features one by one based on the model's feedback, removing those which negatively impact the model's performance the most.

The process is recursive because it retrains the model after each feature removal, ranking the remaining features and removing the least useful again, and repeating the process until the optimal number of features is reached.

RFE with cross-validation helps identify the most important features for a model by testing different subsets and evaluating their performance. It ensures the chosen features improve predictive accuracy by selecting the optimal subset through multiple iterations.

The following is the result we got:

Recursive feature selection

Outer resampling method: Cross-Validated (2 fold)

Resampling performance over subset size:

(Tabulated on the next page)

The top 5 variables (out of 12): Hip girth, Knee girth, Waist girth, Height, Thigh girth

We performed multiple regression on these five variables.

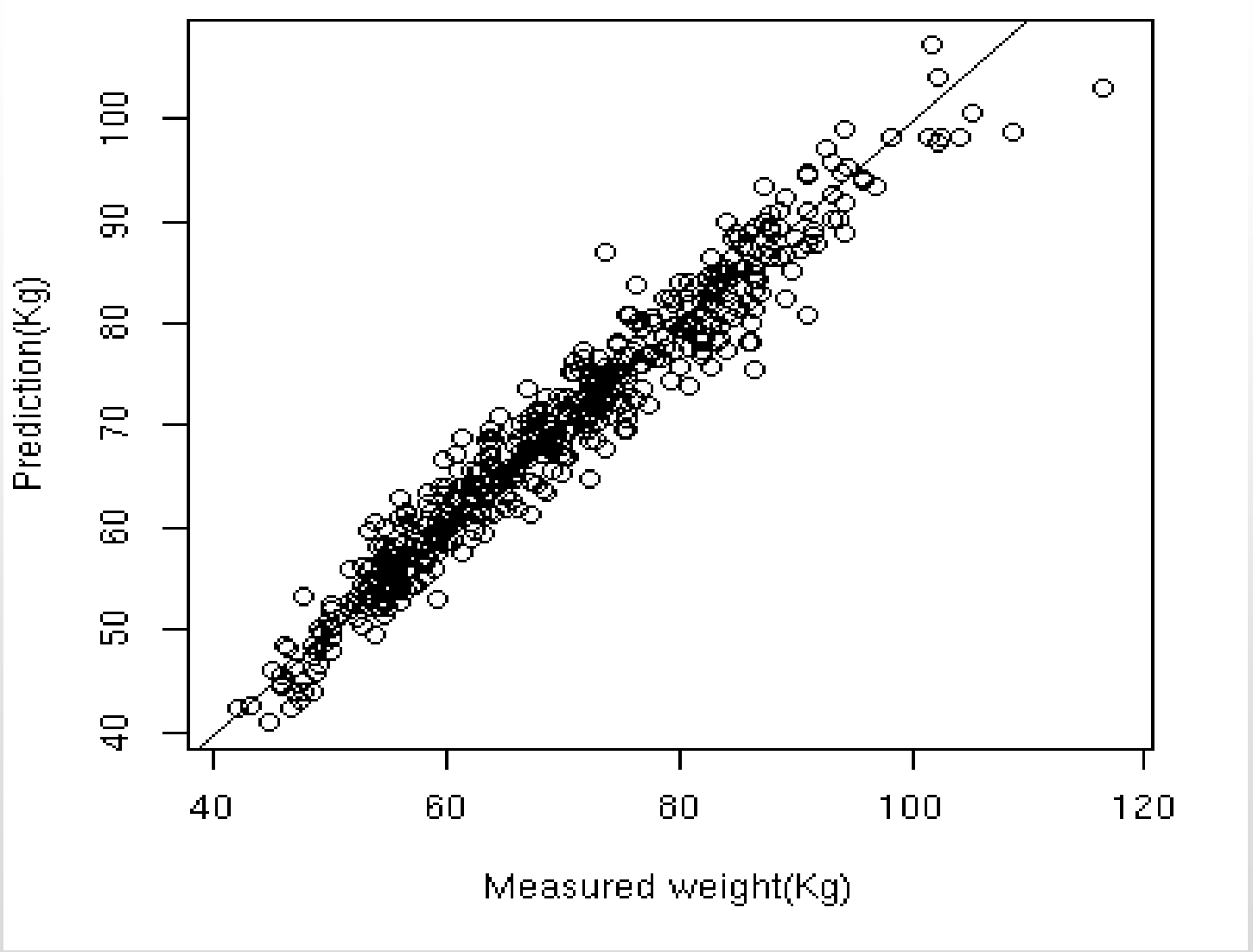
Variables	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
1	9.461	0.5211	7.548	0.46541	0.0155525	0.6700
2	6.350	0.7594	4.989	1.78570	0.1471035	1.5906
3	5.212	0.8494	3.989	0.33020	0.0334514	0.2459
4	4.016	0.9154	3.011	0.59858	0.0142308	0.6393
5	3.742	0.9289	2.849	0.98249	0.0288016	0.8969
6	3.242	0.9437	2.438	0.41102	0.0088892	0.4388
7	2.888	0.9569	2.162	0.04287	0.0048457	0.1392
8	2.897	0.9559	2.169	0.07933	0.0035280	0.2142
9	2.892	0.9556	2.175	0.10739	0.0018714	0.2158
10	2.869	0.9561	2.128	0.11154	0.0018694	0.2423
11	2.888	0.9560	2.170	0.20658	0.0010951	0.2869
12	2.868	0.9566	2.167	0.23221	0.0023316	0.2821
13	2.899	0.9558	2.219	0.19357	0.0013018	0.2808
14	2.876	0.9569	2.192	0.16832	0.0008033	0.2421
15	2.936	0.9548	2.254	0.22804	0.0026165	0.3051
16	2.905	0.9562	2.205	0.21408	0.0022901	0.2859
17	2.932	0.9554	2.240	0.21594	0.0025481	0.2695
18	2.933	0.9548	2.224	0.22826	0.0028830	0.2860
19	2.943	0.9552	2.232	0.23773	0.0027347	0.3167
20	2.939	0.9553	2.227	0.24836	0.0032906	0.3186
21	2.953	0.9551	2.231	0.23742	0.0027677	0.3092
22	2.939	0.9553	2.223	0.24669	0.0030733	0.3149
23	2.950	0.9554	2.232	0.24827	0.0037409	0.3072

Multiple Regression

The Equation we got after performing multiple regression on the aforementioned variables is as follows:-

$$\begin{aligned}\text{Weight} = & -119.55081727 \\ & + 0.66009201 \cdot \text{Waist Girth} \\ & + 0.06031474 \cdot \text{Hip Girth} \\ & + 0.54758229 \cdot \text{Thigh Girth} \\ & + 0.67516796 \cdot \text{Knee Girth} \\ & + 0.44685940 \cdot \text{Height}\end{aligned}$$

Feature selected regression



(Intercept)	bdat[, 14]	bdat[, 12]	bdat[, 18]	bdat[, 24]	bdat[, 15]
-119.55081727	0.06031474	0.66009201	0.67516796	0.44685940	0.54758229

This equation gives us the following values:-

Mean Squared Error (MSE): 9.08

Mean Absolute Error (MAE): 2.27 kg

R^2 Score: 0.949

This model, while performing worse than our previous ones, still performs well while using way less variables. It however functions faster and will be more consistent over new data compared to a model with 23 variables.

Final Inferences

Through our project, we were able to verify the author's claim that the multivariable linear regression model is a better model than the single variable linear regression model. To conclude this, we compared the R^2 scores. We also found that including age as a variable in the regression equation slightly increases the accuracy. This makes sense, as age is an important factor in determining weight. We also compared the CVs obtained from the data set with "standard CVs" and found that the CVs were, in general, higher. This is due to the relatively small sample size and also due to the fact that the sample is not a good representative of the general adult population. We also noticed that for the most part, the body dimensions of men exceed that of women's, but this difference ranges from being negligible to quite apparent.

Acknowledgment

We would firstly like to thank our respected Statistics 1 professor, Dr. Rituparna Sen for giving us the opportunity to work on this project. It was a great learning experience and we got hands-on practice with statistical data analysis.

Thank You