Final Report

Introduction

The dataset we are using was provided by Dr. A. Garth Fisher (personal communication, October 5, 1994), and contains age, weight, height, and body circumference measurements for 252 men. The data was originally collected in 1985 (Penrose et al., 1985) and can be found at Body Fat Prediction Dataset. Instead of using expensive medical imaging methods, these measures were taken in order to create a predictive equation for body fat percentage using simple indicators. The response variable is Body Fat Percentage, estimated by underwater weighing, which is an accurate technique that derives body density and subsequently estimates body fat percentage using predictive equations like Siri's and Brozek's equations. The dataset contains the following variables:

- Bodyfat: Estimated using Siri's and Brozek's equations
- Age: Recorded in years.
- Weight: Recorded in pounds. It includes both weight in air (used in standard conditions) and in water (for density calculation purposes).
- Height: Measured in inches, included in the dataset as a crucial factor for calculating indices like Body Mass Index (BMI).
- Density: Calculated in g/cm^3.
- Body Circumference Measurements: Ten different variables representing various circumferences of Neck, Chest, Abdomen (measured at the umbilicus level with the iliac crest), Hip, Thigh, Knee, Ankle, Biceps (extended), Forearm, Wrist (measured distal to the styloid processes). All measurements are in centimetres (cm) using a measuring tape.
- Density: Body Density in g/cm^3.

Motivation and Research Question

BMI (weight divided by height squared) is widely used as a standard measure of obesity. However, according to Adab et al. (2018), it has limitations. Studies had shown that BMI is not an accurate predictor of obesity, especially in people who are athletic, short or older. Instead, directly measuring body fat distribution appears to be a better indicator of obesity. While technologies like dual-energy X-ray absorptiometry (DEXA), computed tomography (CT), and magnetic resonance imaging (MRI) provide accurate imaging of body fat, they are also expensive and not suitable for repeated use (Denton & Karpe, 2016). This study aimed to evaluate whether simple and cost-effective measurements of body circumference could be a reliable predictor of body fat percentage.

Our research question is **"Which physical measurements of a person, out of the ones we have access to, are related to total body fat percentage?"** The response variable in our investigation will be body fat percentage (from Siri's (1956) equation) and the covariates will be Age, Weight, Height, Neck, Chest, Abdomen, Hip, Thigh, Knee, Ankle, Biceps, Forearm, Wrist. We will omit density as a covariate as it was directly used to calculate body fat percentage using equations and would thus have a direct correlation.

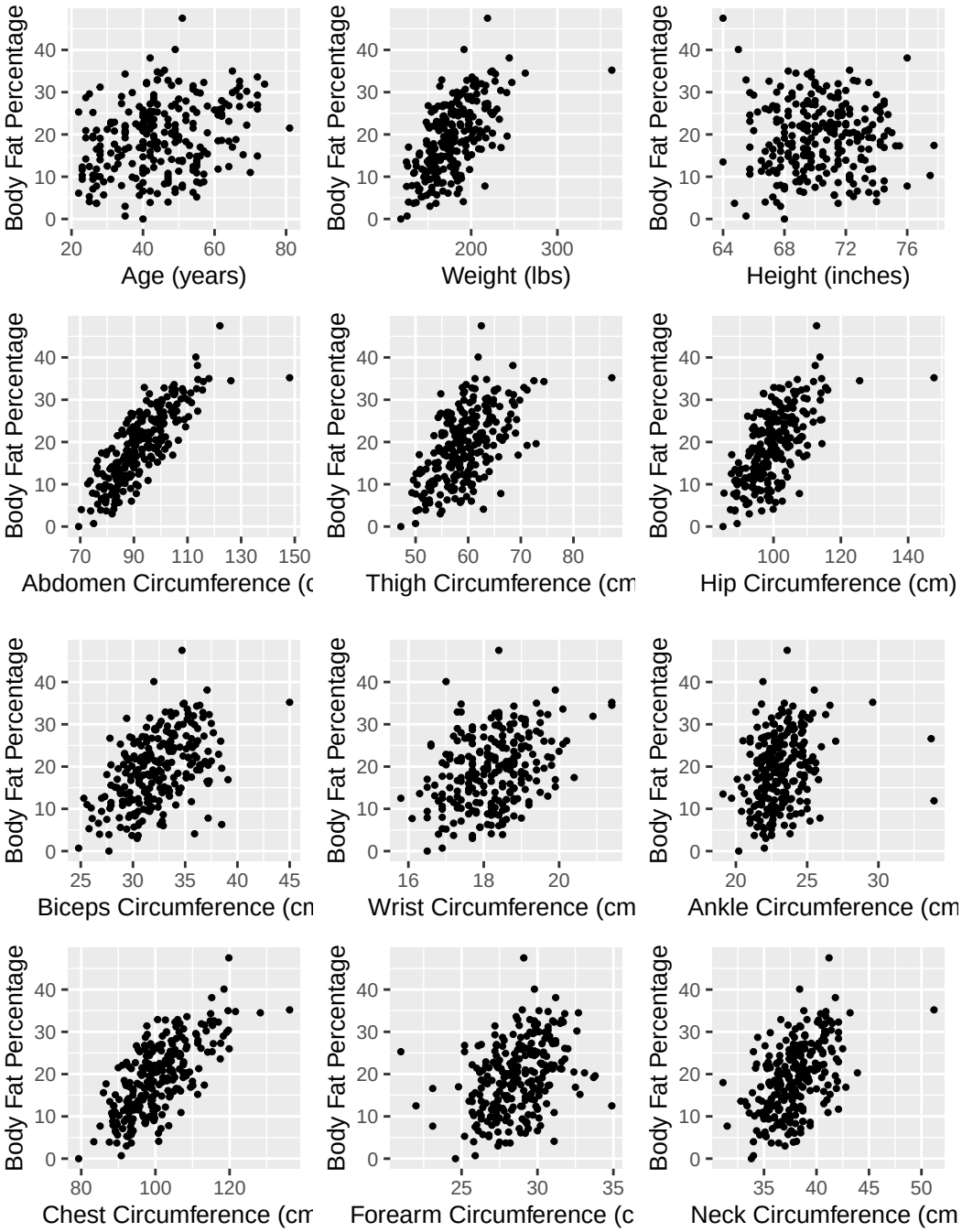
Exploratory Data Analysis

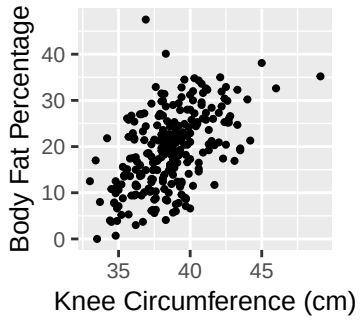
To get an idea of our dataset:

```
# A tibble: 6 x 15
  Density BodyFat   Age Weight Height  Neck Chest Abdomen   Hip Thigh  Knee
    <dbl>   <dbl> <dbl>  <dbl>  <dbl> <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl>
1    1.07    12.3    23   154.   67.8  36.2  93.1    85.2  94.5  59    37.3
2    1.09     6.1    22   173.   72.2  38.5  93.6    83    98.7  58.7  37.3
```

```
3    1.04    25.3    22   154    66.2  34    95.8    87.9  99.2  59.6  38.9
4    1.08    10.4    26   185.   72.2  37.4 102.     86.4 101.   60.1  37.3
5    1.03    28.7    24   184.   71.2  34.4  97.3   100   102.   63.2  42.2
6    1.05    20.9    24   210.   74.8  39   104.     94.4 108.   66    42
# i 4 more variables: Ankle <dbl>, Biceps <dbl>, Forearm <dbl>, Wrist <dbl>
```

Next, we plot body fat percentage against different variables. We remove a data point with Height < 30 from the visualization to better observe the overall trend.





We see that Weight, Abdomen, Thigh, Hip, Biceps, Neck, Ankle, Chest, have moderate to strong positive relationships with Body Fat Percentage. Age, Forearm, Biceps, and Wrist Circumference may have a weak but slightly positive relationship. Height does not seem to have a relationship with Body Fat Percentage.

It does not seem that our variance in bodyfat changes with our covariates, so we don't think a transformation would be needed at this stage of our analysis.

We will also calculate some summary statistics to get an idea of the variables in our dataset.

```
[1] "Means"
```

```
# A tibble: 1 x 14
  BodyFat   Age Weight Height  Neck Chest Abdomen   Hip Thigh  Knee Ankle Biceps
    <dbl> <dbl>  <dbl>  <dbl> <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>
1    19.2  44.9   179.   70.1  38.0  101.    92.6  99.9  59.4  38.6  23.1   32.3
# i 2 more variables: Forearm <dbl>, Wrist <dbl>
```

The average age of a person in our dataset is 44 years. The average bodyfat percentage is 19%. From the sample means, it seems like abdomen, chest and hip circumferences tend to be higher compared to other measurements.

```
[1] "Standard deviations"
```

```
# A tibble: 1 x 14
  BodyFat   Age Weight Height  Neck Chest Abdomen   Hip Thigh  Knee Ankle Biceps
    <dbl> <dbl>  <dbl>  <dbl> <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>
1    8.37  12.6   29.4   3.66  2.43  8.43    10.8  7.16  5.25  2.41  1.69   3.02
# i 2 more variables: Forearm <dbl>, Wrist <dbl>
```

```
# A tibble: 1 x 1
  NA_values
      <int>
1         0
```

There are no NA values in our dataset, which is good for our modelling.

We will now proceed with our regression analysis.

Analysis

Our exploratory data analysis suggested that a linear model may be appropriate to explain the relationship between body fat percentage and various physical measurements. To address our research question, we would attempt to fit a linear regression model, with body fat percentage as our response and a suitable combination of other variables. We would perform backward selection to determine which variables contain the most useful information to explain the variation in body fat percentage.

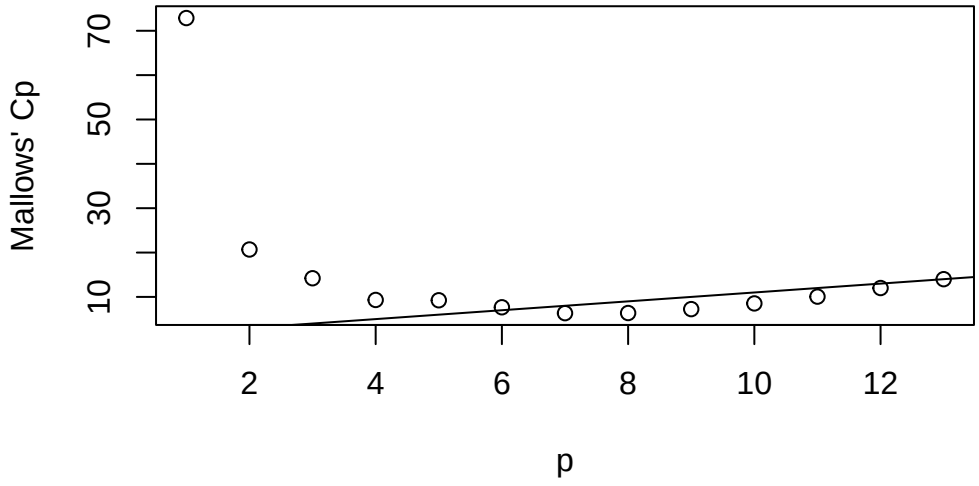
To perform linear regression we need to keep in mind the following assumptions:

1. Linear Relationship between the response and covariates.
2. Independence of error terms
3. Constant Variance of the error terms
4. Normal distribution of error terms

From the exploratory analysis, the linear relationship assumption seems appropriate since most covariates seem to have linear relationship with body fat percentage. While we are not aware of the data collection technique, the data comes from 252 men, so it is reasonable to assume the measurements are independent. Additionally, from our plots, it doesn't seem like the constant variance of error term would be violated. We will later examine these assumptions through appropriate plots.

So now, we perform backwards selection to identify potential models.

To further narrow down our options, we can compute Mallows' C_p statistic for each model, treating the model with all 13 covariates as our full model. We also tabulate the values for R^2 and Adjusted R^2.




```
# A tibble: 13 x 4
       p    Cp    R2 AdjR2
   <int> <dbl> <dbl> <dbl>
 1     1 72.9  0.662 0.660
 2     2 20.7  0.719 0.717
```

```
 3     3 14.2  0.728 0.724
 4     4  9.31 0.735 0.731
 5     5  9.24 0.737 0.732
 6     6  7.66 0.741 0.735
 7     7  6.34 0.744 0.737
 8     8  6.37 0.747 0.738
 9     9  7.25 0.748 0.738
10    10  8.53 0.748 0.738
11    11 10.1  0.749 0.737
12    12 12.0  0.749 0.736
13    13 14.0  0.749 0.735
```

From the above plot of C_p vs p, we see that the only models with C_p values close to the $p + 1$ line are those with 6, 7, 11, 12, and 13 covariates. To decide between these models, we can look at the R^2 and adjusted R^2 values, and we see their rate of increase decreases after $p = 6$, indicating that the benefit of adding additional covariates is smaller. Therefore, we choose the model with 6 covariates: Abdomen, Weight, Wrist, Forearm, Age, and Thigh.

```
Call:
lm(formula = BodyFat ~ Abdomen + Weight + Wrist + Forearm + Age +
    Thigh, data = bodyfat)

Residuals:
     Min       1Q   Median       3Q      Max
-10.8702  -3.0465  -0.1963   3.0774   8.9299

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -38.32154    8.61242  -4.450 1.31e-05 ***
Abdomen       0.91179    0.06975  13.072  < 2e-16 ***
Weight       -0.13648    0.03288  -4.150 4.59e-05 ***
Wrist        -1.77884    0.49469  -3.596 0.000391 ***
Forearm       0.48913    0.18232   2.683 0.007797 **
Age           0.06290    0.03080   2.042 0.042220 *
Thigh         0.22024    0.11656   1.889 0.060009 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.311 on 245 degrees of freedom
Multiple R-squared:  0.741, Adjusted R-squared:  0.7346
F-statistic: 116.8 on 6 and 245 DF,  p-value: < 2.2e-16
```

We see that all of our variables except Thigh are significant at the 5% level. The p-value for thigh is very close to statistical significance. As expected, most variables have a positive coefficient, suggesting that the response (bodyfat percentage) increases with an increase in the covariates.

On initial thought, it is slightly surprising that weight and wrist have a negative coefficient because the exploratory data analysis demonstrated that these variables have a positive relationship with bodyfat percentage.

We can examine this further using partial correlations. We calculate partial correlation between weight and bodyfat while adjusting for different combinations of covariates.

```
Partial correlation between Weight and Bodyfat, adjusting for Abdomen, Wrist,
Forearm, Age, Thigh: -0.256
```

```
Partial correlation between Weight and bodyfat, adjusting for Abdomen: -0.411

Partial correlation between Weight and bodyfat, adjusting for Wrist, Forearm,
Age, Thigh: 0.315

Partial correlation between Weight and bodyfat, adjusting for Forearm, Age,
Thigh: 0.166
```

From the partial correlations, it seems like once we have adjusted for abdomen circumference, a one unit increase in weight is related to a decrease in body fat percentage. This could be because keeping abdomen circumference constant, an increase in weight may be related to muscle gain or increase in bone density. If we did not adjust for abdomen, we see that an increase is weight is associated with increases in body fat, which suggests that body fat in men is possibly stored in abdomen.

Let's do the same for wrist:

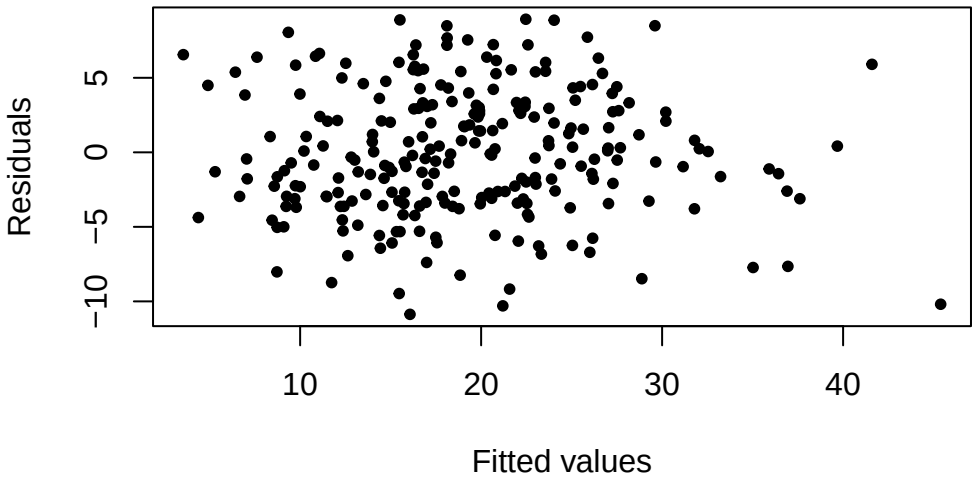
```
Partial correlation between wrist and bodyfat, adjusting for Abdomen, Weight,
Forearm, Age, Thigh: -0.224

Partial correlation between wrist and bodyfat, adjusting for Weight, Forearm, Age,
Thigh: -0.351

Partial correlation between wrist and bodyfat, adjusting for Forearm + Age +
Thigh: -0.232
```

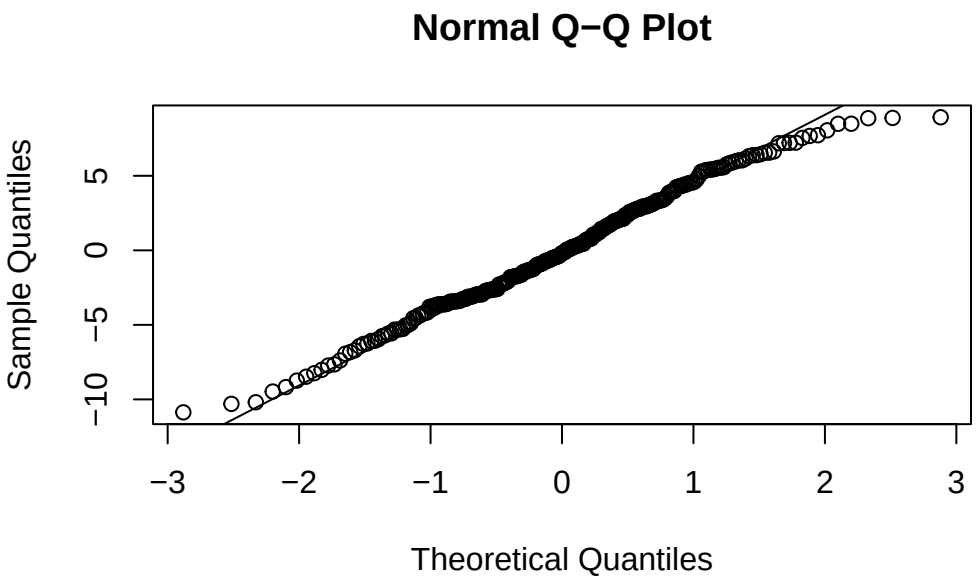
It seems like once we have adjusted for other covariates, wrist is negatively correlated with body fat. It is unclear why this may be the case but we can hypothesize that the same reasoning for the effect of weight can be applied here. It may be that keeping other measurements constant, increase in wrists sizes are due to other body changes indicative of fat loss, such as muscle or bone density gain.

Both these findings suggest that solely wrist and weight may not be good indicators of body fat.

Diagnostics

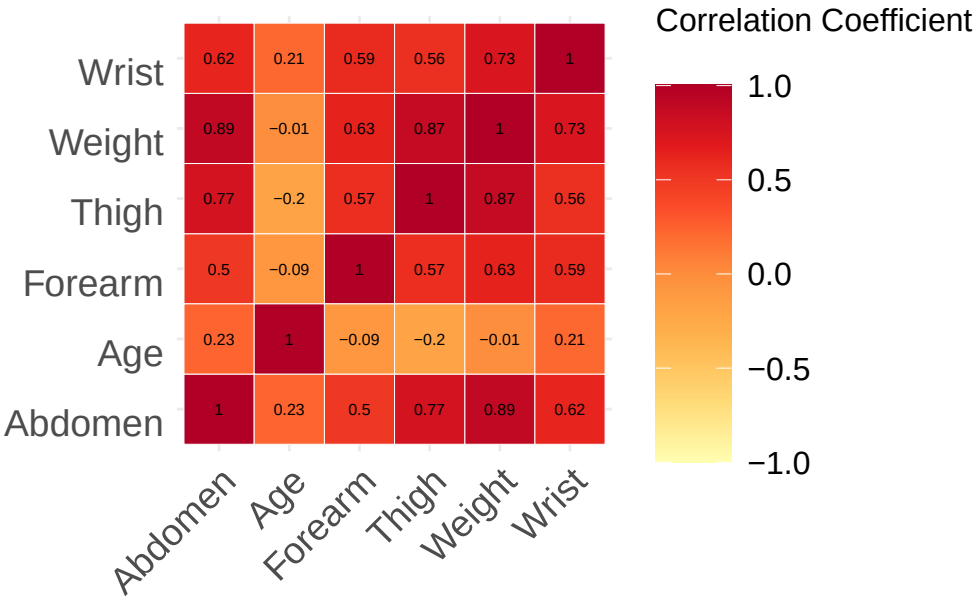
We check residuals plots to ensure assumptions about our model have not been violated:





We see no obvious non-linear or fan-shaped pattern in the residuals vs. fitted values plot, indicating that linearity and homoscedasticity assumptions have not been violated. The QQ plot shows some signs of the errors being light-tailed, but the deviations should be slight enough that our model is a good fit overall.

We can also assess any multicollinierity between input variables, since we can expect some physical measurements to be correlated with one another. We will plot the correlations between the different variables as a heat-map.



This plot shows that the weight variable strongly correlated with thigh and abdomen and moderately correlated with wrist. Abdomen and thigh also seem to be moderately correlated. To further investigate this issue of multicollinearity we can calculate the Variance Inflation Factors (VIFs).

```
  Abdomen    Weight     Wrist   Forearm       Age     Thigh
7.639737 12.613499  2.880501  1.832929  2.035071  5.057444
```

Conclusion

From the analysis, we identified that several physical measurements are significantly related to body fat percentage. After conducting exploratory data analysis and using backward selection, the optimal linear regression model included six covariates: Abdomen, Weight, Wrist, Forearm, Age, and Thigh.

This model was chosen based on its adherence to the assumptions of linear regression, minimal multicollinearity among variables, and strong statistical metrics, including Mallows' Cp, R^2 and adjusted R^2. These 6 covariates seem to be good explanatory variables for bodyfat percentage and while including other variables improves our model, the additional improvement is quite small.

The selected model performed well under diagnostic checks, with residual plots indicating no major violations of linearity, independence, or constant variance assumptions. While the QQ plot showed slight deviations, these were minor and did not significantly undermine the fit of the model. There was some evidence of multicollinearity, which suggests that a smaller subset of our variables could also create a good model, but for the purposes of our analysis, we decided to include some additional variables since they improved adjusted R^2 and did not have high standard errors.

Key findings

- **Abdomen circumference** is the strongest predictor of body fat, as seen by its high coefficient and significance.
- Other variables like **Weight, Wrist, Forearm, Thigh, Age** also contribute meaningfully.
- The model fits the data well but has some residual error (4.311), likely from unmeasured factors or noise.

Implications

The findings suggest that these easily measurable physical attributes can effectively estimate body fat percentage. Abdomen circumference, in particular, stands out as a critical predictor, which could be useful for health and fitness evaluations without requiring sophisticated equipment.

Limitations

The data in our study only comes from men, so the findings may not be applicable to other genders. More data on other genders would be critical in conducting research on the relationship between body fat and physical measurements for these groups.

Additionally, we utilised backward selection to pick an optimal model of 6 covariates. However, backward selection is not exhaustive and can miss potentially useful models. Further research utilising other kinds of model selection techniques like forward selection and best subset selection can shed light on other good indicators of body fat.

Future Directions

While the current model provides meaningful insights, its applicability is limited to the specific population (252 males) studied. Future research should aim to validate these findings across diverse groups, including females and individuals from various ethnic backgrounds and age ranges. Additionally, incorporating factors like diet, activity level, and genetic predisposition could further refine predictions and broaden the scope of applicability.