Final Report

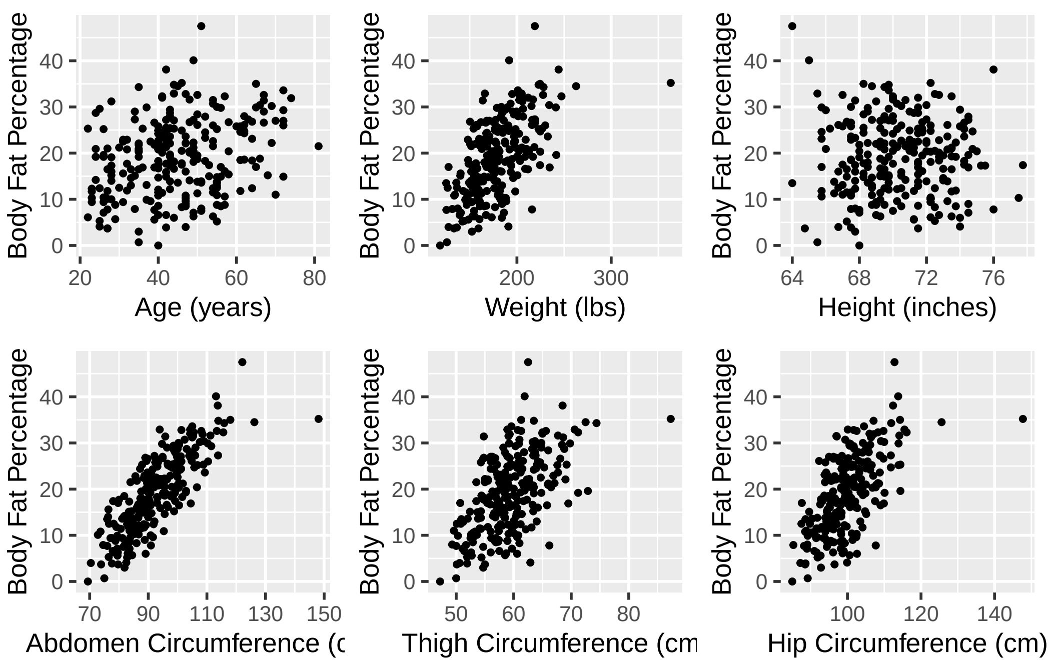
Introduction

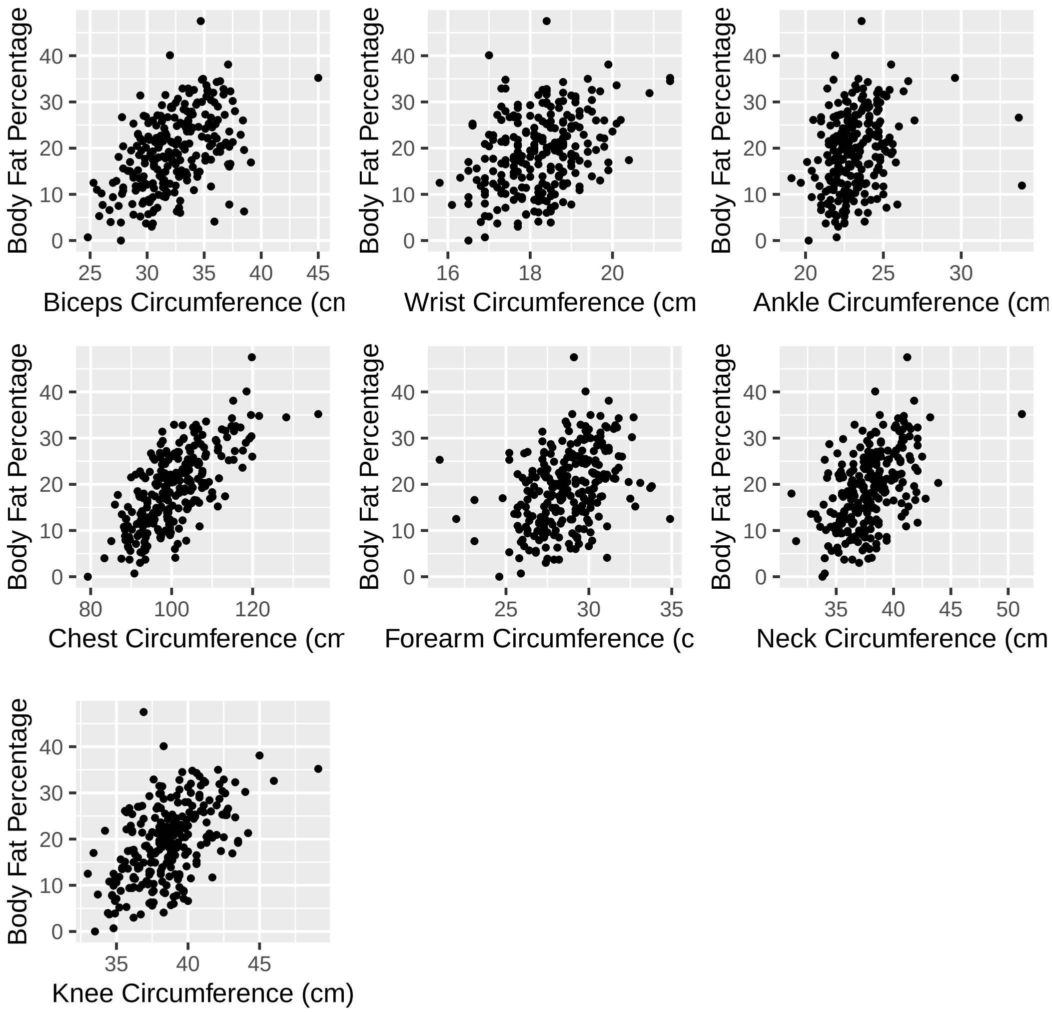
Exploratory Data Analysis

Before using the dataset, we remove density as it was used to calculate body fat percentage and would thus have a direct correlation.

```
# 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 Circumference and Wrist Circumference may have a very 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>
```

```
[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.

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:

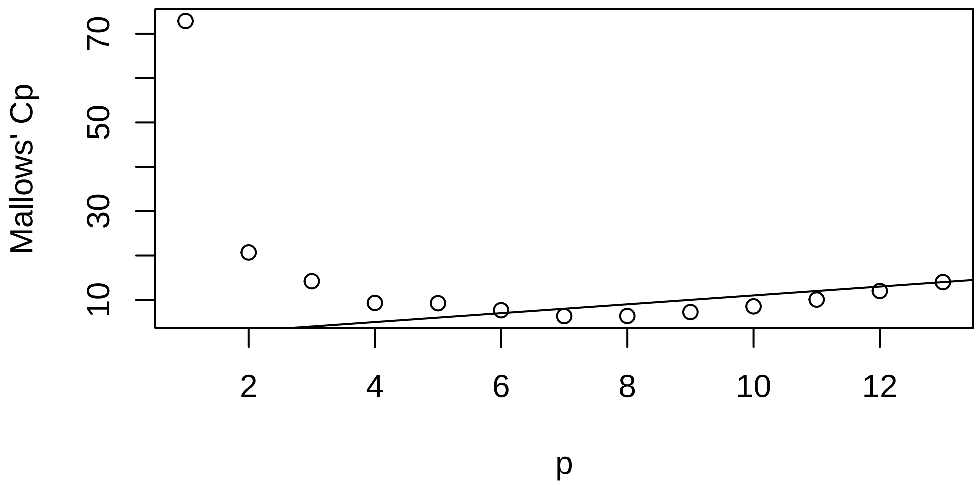
1. Linear Relationship between the response and covariates.
2. Independence of error terms
3. Constant Variance of the error terms
4. Normal distribution of error terms

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.

```
   (Intercept)   Age Weight Height  Neck Chest Abdomen   Hip Thigh  Knee Ankle
1         TRUE FALSE  FALSE  FALSE FALSE FALSE    TRUE FALSE FALSE FALSE FALSE
2         TRUE FALSE   TRUE  FALSE FALSE FALSE    TRUE FALSE FALSE FALSE FALSE
3         TRUE FALSE   TRUE  FALSE FALSE FALSE    TRUE FALSE FALSE FALSE FALSE
4         TRUE FALSE   TRUE  FALSE FALSE FALSE    TRUE FALSE FALSE FALSE FALSE
5         TRUE  TRUE   TRUE  FALSE FALSE FALSE    TRUE FALSE FALSE FALSE FALSE
6         TRUE  TRUE   TRUE  FALSE FALSE FALSE    TRUE FALSE  TRUE FALSE FALSE
7         TRUE  TRUE   TRUE  FALSE  TRUE FALSE    TRUE FALSE  TRUE FALSE FALSE
8         TRUE  TRUE   TRUE  FALSE  TRUE FALSE    TRUE  TRUE  TRUE FALSE FALSE
9         TRUE  TRUE   TRUE  FALSE  TRUE FALSE    TRUE  TRUE  TRUE FALSE FALSE
10        TRUE  TRUE   TRUE  FALSE  TRUE FALSE    TRUE  TRUE  TRUE FALSE  TRUE
11        TRUE  TRUE   TRUE   TRUE  TRUE FALSE    TRUE  TRUE  TRUE FALSE  TRUE
12        TRUE  TRUE   TRUE   TRUE  TRUE  TRUE    TRUE  TRUE  TRUE FALSE  TRUE
13        TRUE  TRUE   TRUE   TRUE  TRUE  TRUE    TRUE  TRUE  TRUE  TRUE  TRUE
   Biceps Forearm Wrist
1   FALSE   FALSE FALSE
2   FALSE   FALSE FALSE
3   FALSE   FALSE  TRUE
4   FALSE    TRUE  TRUE
5   FALSE    TRUE  TRUE
6   FALSE    TRUE  TRUE
7   FALSE    TRUE  TRUE
8   FALSE    TRUE  TRUE
9    TRUE    TRUE  TRUE
10   TRUE    TRUE  TRUE
11   TRUE    TRUE  TRUE
12   TRUE    TRUE  TRUE
13   TRUE    TRUE  TRUE
```

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
```

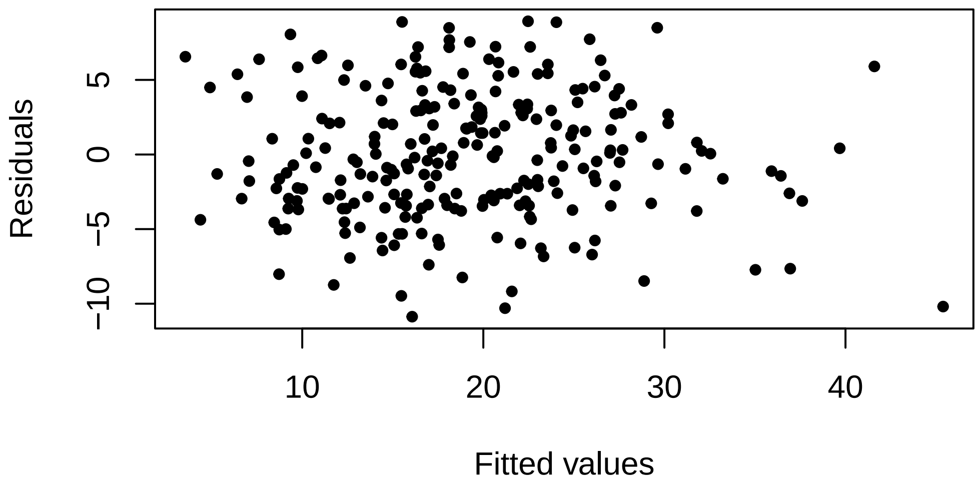
We check residuals plots to ensure assumptions about our model have not been violated:

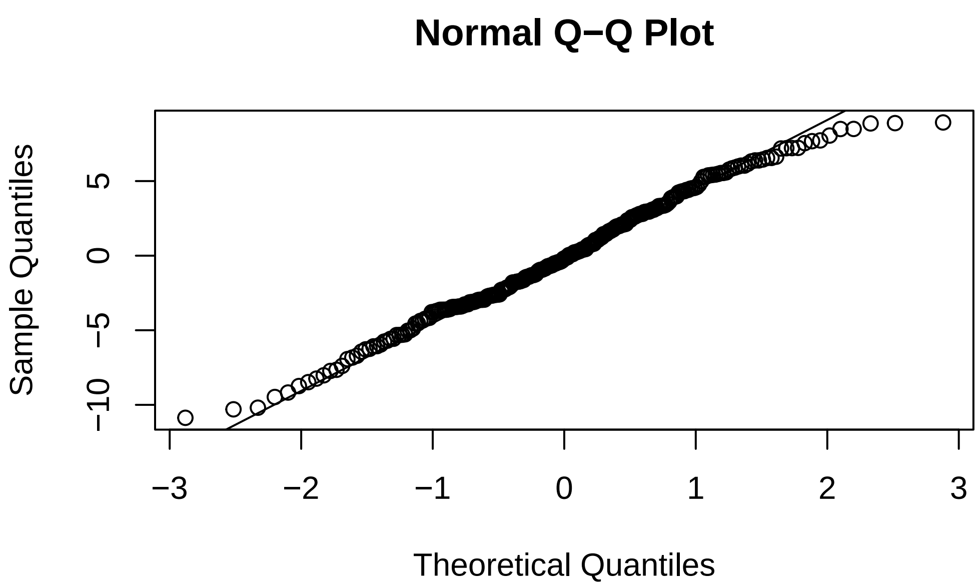

```
Call:
lm(formula = BodyFat ~ Abdomen + Weight + Wrist + Forearm + Age +
    Thigh, data = bodyfat)

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 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.

Conclusion