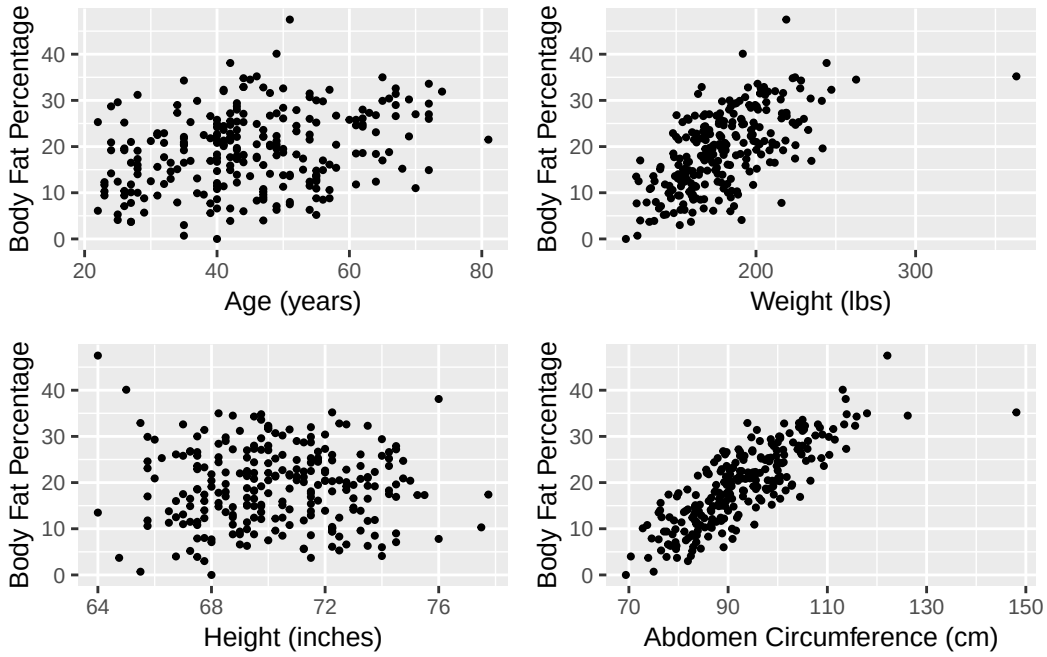
Final Report

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.

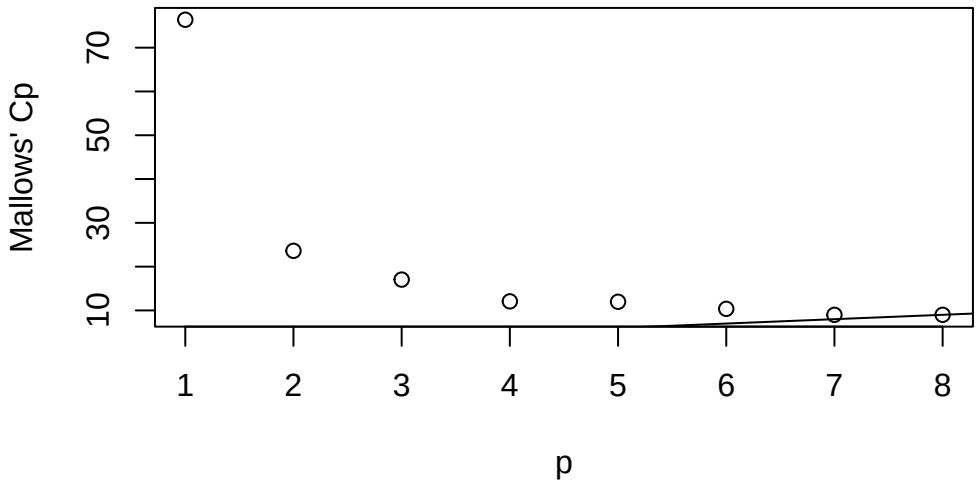


We see that Weight and Abdomen Circumference have moderate to strong positive relationships with Body Fat Percentage, while Age may have a very weak but slightly positive relationship. Height does not seem to have a relationship with Body Fat Percentage.

Next, 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
  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
```

To further narrow down our options, we can compute Mallows' C_p statistic for each model, treating the model with 8 covariates as our full model.

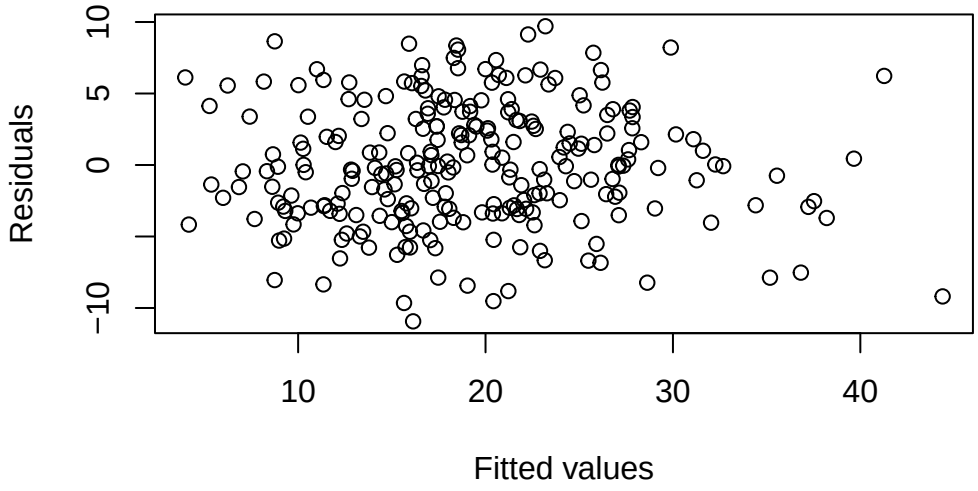


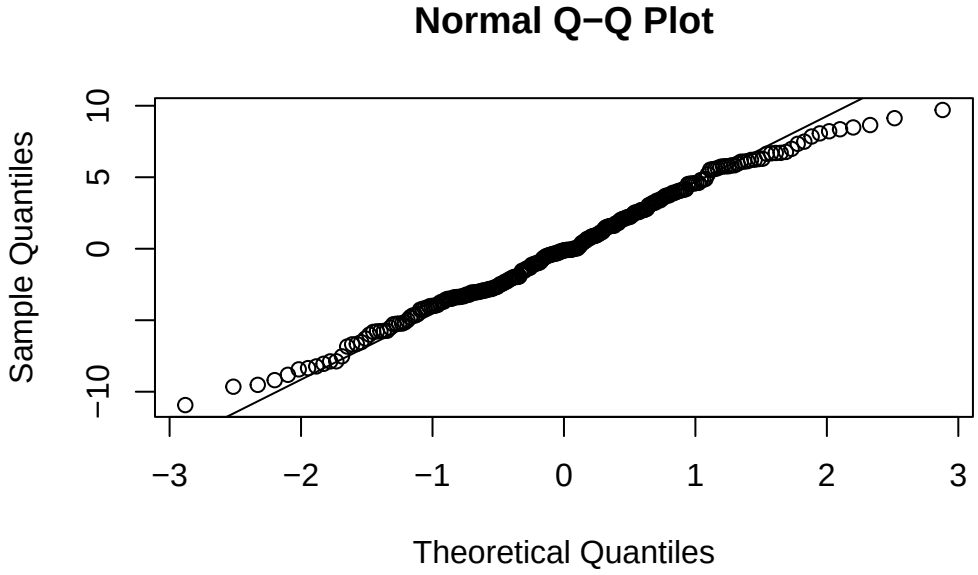
From the above plot of C_p vs p, we see that the only models with C_p values close to the $p+1$ line are those with 7 and 8 covariates. Let us compare them below.

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

Residuals:
    Min      1Q  Median      3Q     Max
-10.936  -3.046  -0.112   3.168   9.705

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -33.25799    9.00681  -3.693 0.000274 ***
Abdomen       0.91788    0.06950  13.207  < 2e-16 ***
Weight       -0.11944    0.03403  -3.510 0.000533 ***
Wrist        -1.53240    0.51041  -3.002 0.002958 **
Forearm       0.55314    0.18479   2.993 0.003043 **
Age           0.06817    0.03079   2.214 0.027769 *
Thigh         0.22196    0.11601   1.913 0.056888 .
```

```
Neck         -0.40380    0.22062  -1.830 0.068424 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.291 on 244 degrees of freedom
Multiple R-squared:  0.7445,    Adjusted R-squared:  0.7371
F-statistic: 101.6 on 7 and 244 DF,  p-value: < 2.2e-16



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

Residuals:
     Min       1Q   Median       3Q      Max
-10.9757  -2.9937  -0.1644   2.9766  10.2244

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -22.65637   11.71385  -1.934  0.05426 .
Abdomen       0.94482    0.07193  13.134  < 2e-16 ***
Weight       -0.08985    0.03991  -2.252  0.02524 *
Wrist        -1.53665    0.50939  -3.017  0.00283 **
Forearm       0.51572    0.18631   2.768  0.00607 **
Age           0.06578    0.03078   2.137  0.03356 *
Thigh         0.30239    0.12904   2.343  0.01992 *
Neck         -0.46656    0.22462  -2.077  0.03884 *
Hip          -0.19543    0.13847  -1.411  0.15940
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.282 on 243 degrees of freedom
Multiple R-squared:  0.7466,    Adjusted R-squared:  0.7382
F-statistic: 89.47 on 8 and 243 DF,  p-value: < 2.2e-16
```

Since the Hip variable added in the model with 8 covariates is not statistically significant, we will choose the model with 7 covariates: Abdomen, Weight, Wrist, Forearm, Age, Thigh, Neck. 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 small enough that our model is a good fit overall.