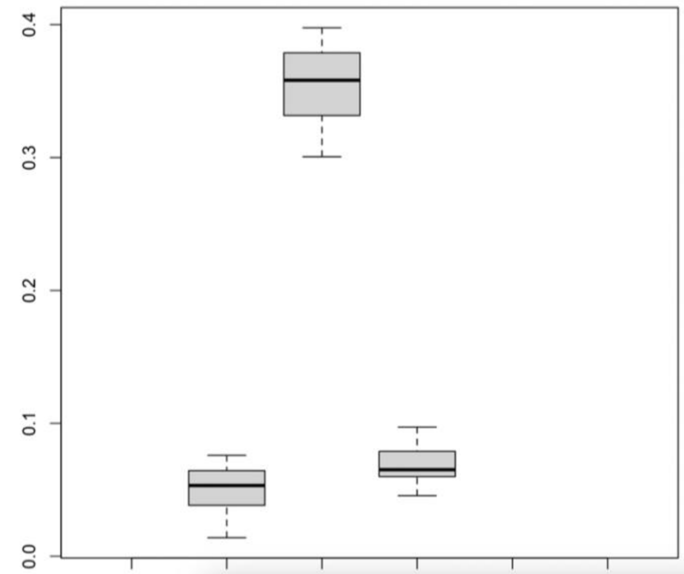
Inefficient Walking Report

To preface, all analysis was completed using R in R Studio. Given our initial data, including measurements of body mass index (BMI) and categorizations of gender for each participant and each walk. One important measure of these walks is their energy expenditure (EE). So, to first find a visual representation of the EE data of all walks, side by side for comparison, is crucial. This is what the following graph displays.



From left to right, the first boxplot summarizes the EE data of the usual walk, the following in the data for the Mr. Teabag walk, and the third is for the Mr. Putey walk. What these boxplots tell us is that, comparatively, the walk performed by the participant mimicking the walk of Mr. Teabag is the most inefficient walk, in terms of energy expenditure. While the other two walks, the usual walk and the walk emulating the walk of Mr. Putey are similar in EE. However, the walk of Mr. Putey is seen to still be less efficient on average than the usual walk. To further analyze the data, there should be interest in the variance between the walks, as quite obviously the EE of all three walks differ. To do so, the Analysis of Variance Table was computed. The table is as follows:

```
Analysis of Variance Table

Response: wvector
          Df  Sum Sq Mean Sq F value    Pr(>F)
walks      2 0.75284 0.37642  672.01 < 2.2e-16 ***
Residuals 36 0.02017 0.00056
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

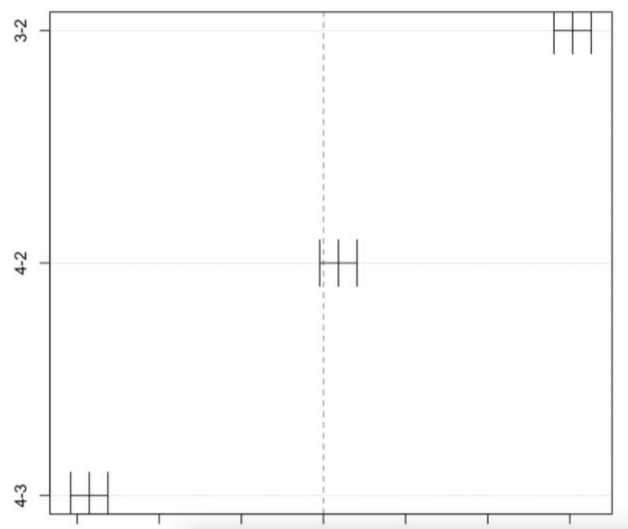
Given this information, we should reject if:

$$672.01 > \mathrm{F}_{1-\alpha}\big(\mathrm{I}-1, \mathrm{I}(\mathrm{J}-1)\big) = \mathrm{F}_{0.95}(2{,}24) = 3.40$$

Thus, we should reject. Next, we perform Tukey's Honest Significance Test to see how significant the difference between specific pairs of walks is. The table given by R following the computation is below:

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = wvector ~ walks, data = newdata)

$walks
           diff          lwr         upr     p adj
3-2  0.30340648  0.280715885  0.32609707 0.0000000
4-2  0.01819472 -0.004495871  0.04088531 0.1369009
4-3 -0.28521176 -0.307902347 -0.26252116 0.0000000
```

Using this data, and a significance level of 0.05, a plot was made:



Again from left to right, the first pair (4-3) is Mr. Teabag and Mr. Putey, then (4-2) which is Mr. Putey and the Usual walk, and finally (3-2) is the Usual walk and the Mr. Teabag walk. The next measure we want to take is to perform a multiple linear regression on Mr. Teabag's walk with both gender and BMI. This model would utilize the generic model but tailored to the situation would be:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + e_i$$

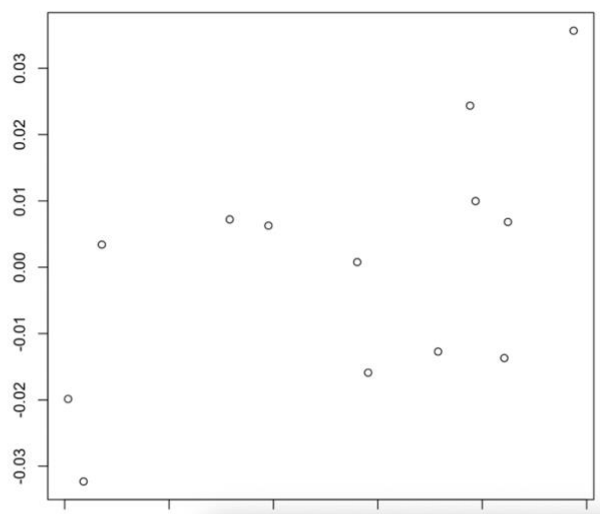
The linear regression was calculated and summarized as below:

```
Call:
lm(formula = Mr.Teabag ~ gender + BMI, data = lrdata)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.032312 -0.013699  0.003413  0.007208  0.035670 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.170981   0.038727   4.415  0.00130 ** 
gender      -0.019904   0.012811  -1.554  0.15133    
BMI          0.007524   0.001601   4.701  0.00084 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02039 on 10 degrees of freedom
Multiple R-squared:  0.6933,  Adjusted R-squared:  0.632 
F-statistic:  11.3 on 2 and 10 DF,  p-value: 0.002712
```

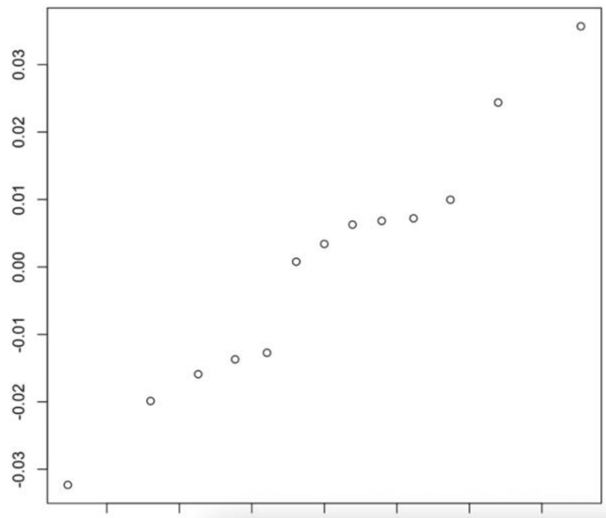
Using this table, it is clear that while BMI did impact our energy efficiency in relation to the Mr. Teabag walk, gender actually did not play a role in difference. Our linear regression equation would then be:

$$Y_i = 0.170981 - 0.19904X_{i,1} + 0.007524X_{i,2} + e_i$$

Finally, we should assess whether or not this model above is appropriate. This is because there are four crucial assumptions that are made when a linear regression model is applied. The assumptions are as follows: linearity of parameters, constant variance, independence of data sets, and normality of error. The following is the plot of Mr. Teabag’s walk versus the residuals.



Although the plot is scattered, it still averages near zero and we can consider this to imply constant variance. To further this conclusion, the Q-Q plot for these residuals is as follows:



While there is slight deviation towards the upper right portion of the graph, we can generally see this to imply normality. Thus, this model is appropriate for this situation as the lack of pattern in the residual plot and the normality of the Q-Q plot are evidence that the model is useful here. Our conclusions about the linear regression parameters then are accurate and help to describe the regression.