BMI Walking Stance Project Final Report

Suheng Yao, Jonathan Neimann, Yishun Zhang

2025-05-02

Abstract

Obesity significantly impacts musculoskeletal health and alters walking biomechanics. This study aimed to quantify the influence of obesity on knee angle changes during various walking tasks and identify associated anthropometric predictors. We analyzed motion sensor data from 35 participants (12 normal-weight, 23 obese) performing six distinct walking tasks (e.g., preferred speed, obstacle negotiation). Knee angle difference (InitialPeak) was the primary outcome. Exploratory data analysis (EDA), including correlation analysis and Variance Inflation Factor (VIF) assessment, guided variable selection. Linear Mixed-Effects Models (LMM), Generalized Additive Mixed Models (GAMM), and Bayesian Mixed Models were developed, incorporating fixed effects for group, demographics, and selected body measurements, with random intercepts for participant and task. Model performance was compared using ANOVA and 5-fold cross-validation (RMSE, R^2). Two-sample t-tests compared knee angle differences between groups for each task. EDA and t-tests revealed significantly reduced knee angle differences in the obese group across most tasks ($p < 0.05$), suggesting less knee flexion. The final LMM demonstrated the best fit and predictive performance (Avg CV RMSE ≈ 2.59, Avg CV $R^2 \approx 0.68$), identifying significant associations between knee angle difference and Shoulder Breadth (positive), Chest Breadth (negative), Lower Thigh Circumference (negative), Shank Circumference (negative), and A Body Shape Index (ABSI, negative). Significant variability was attributed to both participant and task random effects. In conclusion, obesity is associated with reduced knee flexion during walking, and specific body dimensions beyond BMI contribute significantly to these biomechanical alterations.

Introduction

Obesity is a growing public health concern worldwide, with well-documented impacts on musculoskeletal health and mobility. Excess body weight alters biomechanical loading patterns during everyday activities, increasing the risk of joint degeneration and chronic musculoskeletal conditions. In particular, deviations in walking mechanics—such as reduced knee flexion and altered foot–ground contact—have been observed in individuals with obesity, suggesting a potential pathway by which obesity contributes to long-term joint dysfunction and osteoarthritis.

Knee angle during gait is a critical marker of walking stance and limb mechanics. This project aims to quantify how obesity influences knee-angle trajectories across a series of controlled walking conditions. Thirty-five participants (12 normal-weight, 23 obese) underwent six distinct walking tasks—ranging from preferred and fast speeds to obstacle approaches and crossings at two heights—while equipped with motion sensors on the upper leg, lower leg, and shoes. By comparing knee-angle profiles between obese and non-obese groups, we seek to identify the association between knee angles and people's body measurements.

In this report, we will mainly talk about the EDA of the dataset, the models used to assess the relationship and interpret our models' results to reach a final conclusion.

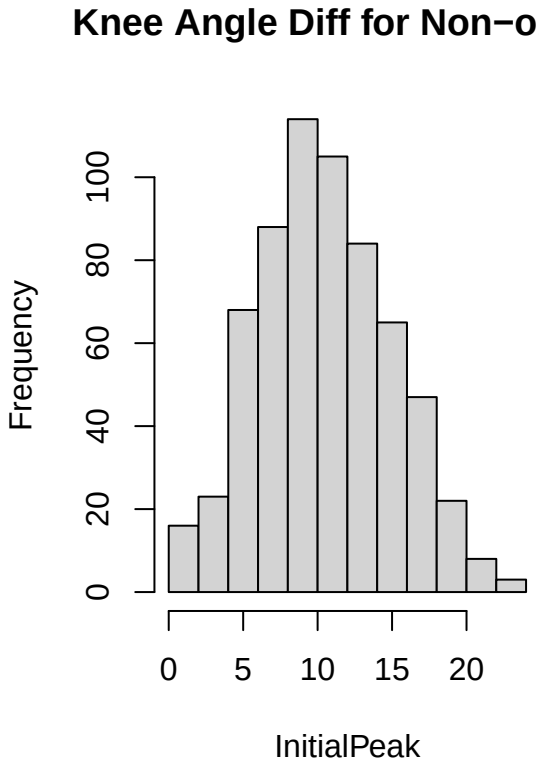
Data Description

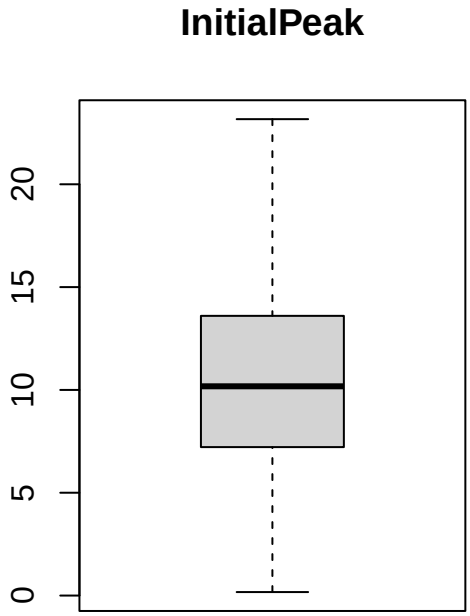
```
## 'data.frame':    1826 obs. of  47 variables:
##  $ N          : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ studyid    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ age        : int  36 36 36 36 36 36 36 36 36 36 ...
##  $ Group      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ BS         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Sex        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ BMI        : num  20.1 20.1 20.1 20.1 20.1 ...
##  $ Race       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ height_m   : num  1.66 1.66 1.66 1.66 1.66 1.66 1.66 1.66 1.66 1.66 ...
##  $ weight_kg  : num  55.3 55.3 55.3 55.3 55.3 ...
##  $ leg_l_l    : num  92 92 92 92 92 92 92 92 92 92 ...
##  $ leg_l_r    : num  93 93 93 93 93 93 93 93 93 93 ...
##  $ leg_l      : num  92.5 92.5 92.5 92.5 92.5 92.5 92.5 92.5 92.5 92.5 ...
##  $ DST        : int  22 22 22 22 22 22 22 22 22 22 ...
##  $ Stroop     : int  119 119 119 119 119 119 119 119 119 119 ...
##  $ PA         : num  7.12 7.12 7.12 7.12 7.12 ...
##  $ Task       : chr  "PRF" "PRF" "PRF" "PRF" ...
##  $ Trial      : int  1 2 3 4 1 2 3 4 1 2 ...
##  $ Speed      : num  1.29 1.32 1.32 1.32 1.55 1.51 1.57 1.53 1.4 1.35 ...
##  $ Initial    : num  4.7 4.06 3.62 4.47 7.74 ...
##  $ Peak       : num  10.4 11.8 10.1 12.8 16.3 ...
##  $ InitialPeak: num  5.7 7.79 6.53 8.31 8.61 ...
##  $ Min        : num  1.4 1.51 1.59 2.34 1.5 ...
##  $ MinPeak    : num  9 10.33 8.56 10.44 14.84 ...
##  $ head_cir   : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ neck_cir   : num  36 36 36 36 36 36 36 36 36 36 ...
##  $ SH_B       : num  34.6 34.6 34.6 34.6 34.6 34.6 34.6 34.6 34.6 34.6 ...
##  $ SH_D       : num  19 19 19 19 19 19 19 19 19 19 ...
##  $ CH_B       : num  26.5 26.5 26.5 26.5 26.5 26.5 26.5 26.5 26.5 26.5 ...
##  $ CH_D       : num  18 18 18 18 18 18 18 18 18 18 ...
##  $ WA_B       : num  23 23 23 23 23 23 23 23 23 23 ...
##  $ WA_D       : num  19 19 19 19 19 19 19 19 19 19 ...
##  $ HIP_B      : num  29.4 29.4 29.4 29.4 29.4 29.4 29.4 29.4 29.4 29.4 ...
##  $ HIP_D      : num  20.5 20.5 20.5 20.5 20.5 20.5 20.5 20.5 20.5 20.5 ...
##  $ ASIS       : num  21.3 21.3 21.3 21.3 21.3 21.3 21.3 21.3 21.3 21.3 ...
##  $ waist_cir  : num  76 76 76 76 76 76 76 76 76 76 ...
##  $ hip_cir    : num  90 90 90 90 90 90 90 90 90 90 ...
##  $ thigh_cir  : num  49 49 49 49 49 49 49 49 49 49 ...
##  $ L_thigh_cir: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ shank_cir  : num  29 29 29 29 29 29 29 29 29 29 ...
##  $ ankle_cir  : num  21 21 21 21 21 21 21 21 21 21 ...
##  $ W.H.ratio  : num  0.844 0.844 0.844 0.844 0.844 ...
##  $ ABSI       : num  79.8 79.8 79.8 79.8 79.8 ...
##  $ Hip.Index  : num  50.2 50.2 50.2 50.2 50.2 ...
##  $ biceps_cir : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ forearm_cir: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ wrist_cir  : num  NA NA NA NA NA NA NA NA NA NA ...
```

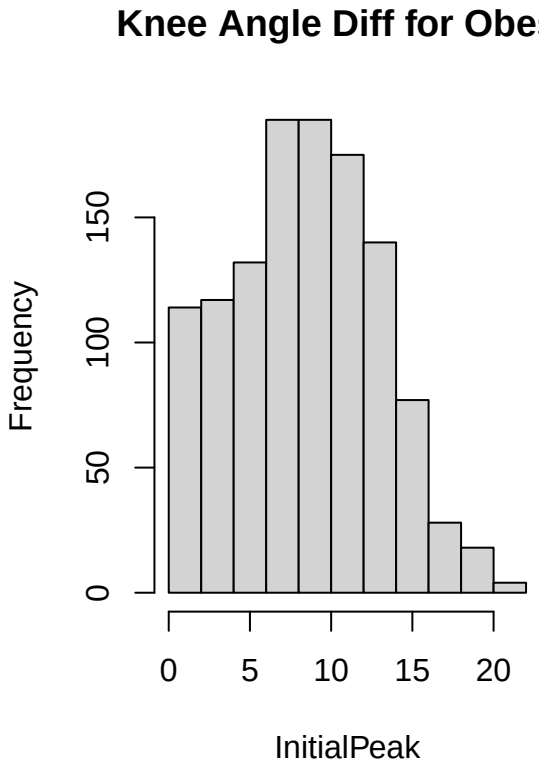
The main dataset used is "BodyShape.csv". It contains the body shape measures, such as BMI, hip circumference, waist hip ratio, cognitive test scores, physical activity scores and the knee angle before and after conducting the task and the difference between those two measures for all 38 participants. However, there

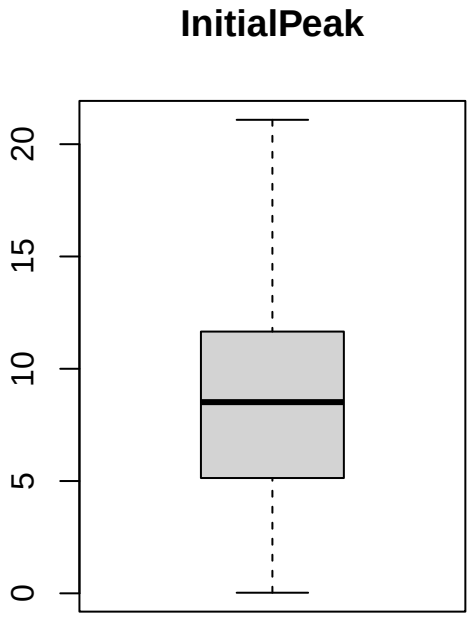
are many repeated measures for each participants, which means that linear mixed effect model would be a good baseline model to start with. Next, we will perform EDA to select the important variables related to knee angles.

EDA

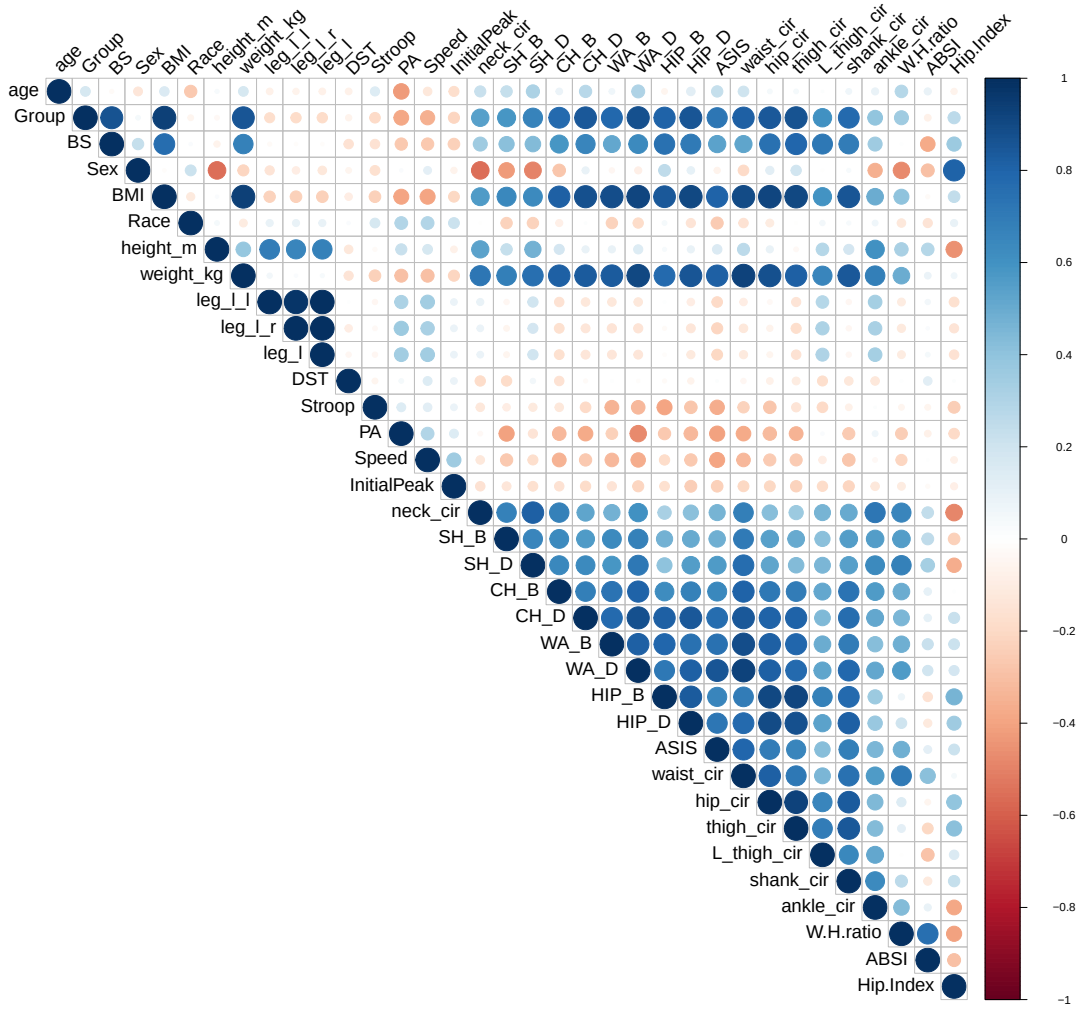

Knee Angle Diff for Non−obese
Frequency
0
20
40
60
80
100
0
5
10
15
20
InitialPeak


InitialPeak
0
5
10
15
20


Knee Angle Diff for Obese
Frequency
0
50
100
150
0
5
10
15
20
InitialPeak


InitialPeak
0
5
10
15
20

Based on the histogram, for normal people, the knee angle difference is mostly centered around 10 degrees, for obese people, the knee angle difference is more right tailed, there are more observations distributed around 0 to 5 degrees; for boxplot, the median of obese people is slightly lower than normal people. Both plots have showed that the knee angle difference is smaller for obese group, meaning they tend to bend less their knees during the six walking stances.



From the correlation plot above, InitialPeak does not have high correlation with any of the body measurements, also, many of the body measurements are highly correlated with each other, and they are also tend to correlate with Group, Body Shape, BMI and weight. We could further use VIF value to test for the multicollinearity.

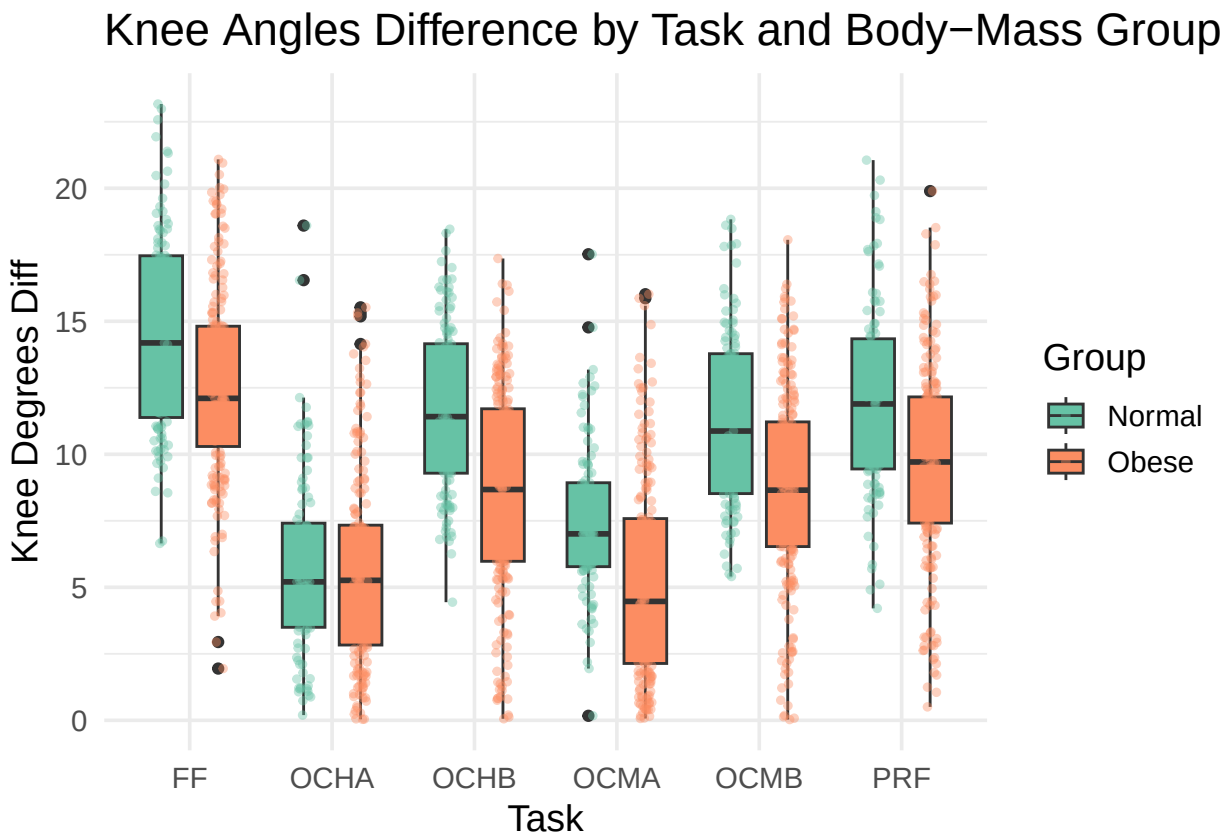
```
##          age        Group           BS          Sex          BMI         Race
##    25.108166   172.590536  162.264620  157.960725 1218.618106   21.039804
##     height_m    weight_kg        leg_l          DST       Stroop           PA
##   341.978116 1330.803293   51.097282    7.665410   19.405663   60.015052
##        Speed     neck_cir         SH_B         SH_D         CH_B         CH_D
```

```
##     1.741173    54.092492    14.205617    46.025297    48.439456    95.524237
##         WA_B         WA_D         HIP_B         HIP_D         ASIS    waist_cir
##   317.478773   291.593649   374.852086    73.216171    94.711758 17359.153040
##      hip_cir    thigh_cir  L_thigh_cir    shank_cir    ankle_cir    W.H.ratio
## 11368.778451   143.166024    24.711372    40.755577    94.422337  8342.737125
##         ABSI    Hip.Index
##   460.558876    33.545880
```

The values above are the values of VIF, and it is clear that most of the body measurement are high correlated with each other since many of the VIF values are far exceeding 10. We will try to remove some of the variables to see if VIF values change.

```
##         age       Group         Sex        Race        leg_l         DST
##    1.891444    8.545979    4.786103    1.884227    2.679177    1.555893
##      Stroop          PA       Speed    neck_cir        SH_B        CH_B
##    2.651510    2.917773    1.627167    7.419055    4.688763    4.606020
##       HIP_B       HIP_D        ASIS L_thigh_cir   shank_cir   ankle_cir
##    7.066339    7.391621    5.102690    6.173321    5.794245    4.829239
##        ABSI
##    2.335713
```

As the results shown above, after removing leg_l_l, leg_l_r, BMI, waist_cir, hip_cir, thigh_cir, height_m, weight_kg, W.H.ratio, Hip.Index, WA_B, WA_D, BS, SH_D and CH_D, most of the VIF become lower than 10, thus, we will try include those variables in the later modeling part.



Based on the boxplot shown above, both groups bent their knees the most when they walk fast, and for all six tasks, people in normal group tend to have greater knee angle difference than the obese group.

Modeling

Model 1: Base Model

```
m1 <- lmer(InitialPeak ~ Group + (1|studyid) + (1|Task), data=df)
```

Model 2: Add Some Demographic Info, Cognitive Test Results and Physical Activity Score

```
m2 <- lmer(InitialPeak ~ Group+age+Sex+Race+leg_l+DST+Stroop+PA+(1|studyid)+(1|Task), data=df)
```

Model 3: Add More Body Measurement Features

```
m3 <- lmer(InitialPeak ~ Group+age+Sex+Race+leg_l+
             DST+Stroop+PA+Speed+neck_cir+SH_B+CH_B+
             HIP_B+HIP_D+ASIS+L_thigh_cir+shank_cir+
             ankle_cir+ABSI+(1|studyid)+(1|Task),
           data=df)
summary(m3)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: InitialPeak ~ Group + age + Sex + Race + leg_l + DST + Stroop +
##     PA + Speed + neck_cir + SH_B + CH_B + HIP_B + HIP_D + ASIS +
##     L_thigh_cir + shank_cir + ankle_cir + ABSI + (1 | studyid) +
##     (1 | Task)
##    Data: df
##
## REML criterion at convergence: 8796.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.9044 -0.6182  0.0022  0.6298  3.8424
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  studyid  (Intercept) 24.101   4.909
##  Task     (Intercept)  7.351   2.711
##  Residual              6.511   2.552
## Number of obs: 1826, groups:  studyid, 25; Task, 6
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept) -11.48042   77.85110  -0.147
## Group         0.88102    1.86180   0.473
## age           0.12140    0.15572   0.780
## Sex          -0.44293    2.65443  -0.167
## Race          0.69062    0.56633   1.219
```

```
## leg_l         0.19914    0.10061   1.979
## DST           0.07188    0.17316   0.415
## Stroop        0.06365    0.60167   0.106
## PA            0.06258    0.73722   0.085
## Speed         1.30061    0.69608   1.868
## neck_cir      0.09850    0.50787   0.194
## SH_B          0.87595    0.17265   5.074
## CH_B         -0.54526    0.21456  -2.541
## HIP_B         0.31224    0.39539   0.790
## HIP_D         0.09220    0.35644   0.259
## ASIS          0.80981    0.43473   1.863
## L_thigh_cir  -0.44396    0.07986  -5.559
## shank_cir    -0.72106    0.29456  -2.448
## ankle_cir     0.41151    0.70122   0.587
## ABSI         -0.36599    0.10188  -3.592
```

Since SH_B(Shoulder Breadth), CH_B(Chest Breadth), L_thigh_cir(lower thigh circumference), shank_cir(shank circumference) and ABSI(a body shape index) are the only statistically significant variables based on the model output above. We can try only including those variables as fixed effects and check the performance using ANOVA later.

Model 4: Only Select Statistically Significant Variables in Model 3

Another variable selection method we could try is Lasso based on the final model:

```
## Optimal lambda (minimum CV error): 0.001778008
```

```
##
## --- Coefficients using lambda.min ---
```

```
## 20 x 1 sparse Matrix of class "dgCMatrix"
##                       s1
## (Intercept) 95.03428556
## Group        3.04010754
## age         -0.01310076
## Sex         -4.01217927
## Race         0.54738873
## leg_l        0.04130643
## DST         -0.09777199
## Stroop      -0.54691509
## PA           0.43883586
## Speed        7.61053976
## neck_cir    -0.59025971
## SH_B         0.16747496
## CH_B         0.17199782
## HIP_B        0.06649541
## HIP_D       -0.36208387
## ASIS         0.10732017
## L_thigh_cir -0.22682117
## shank_cir    0.02821529
## ankle_cir   -0.18040636
## ABSI        -0.11414215
```

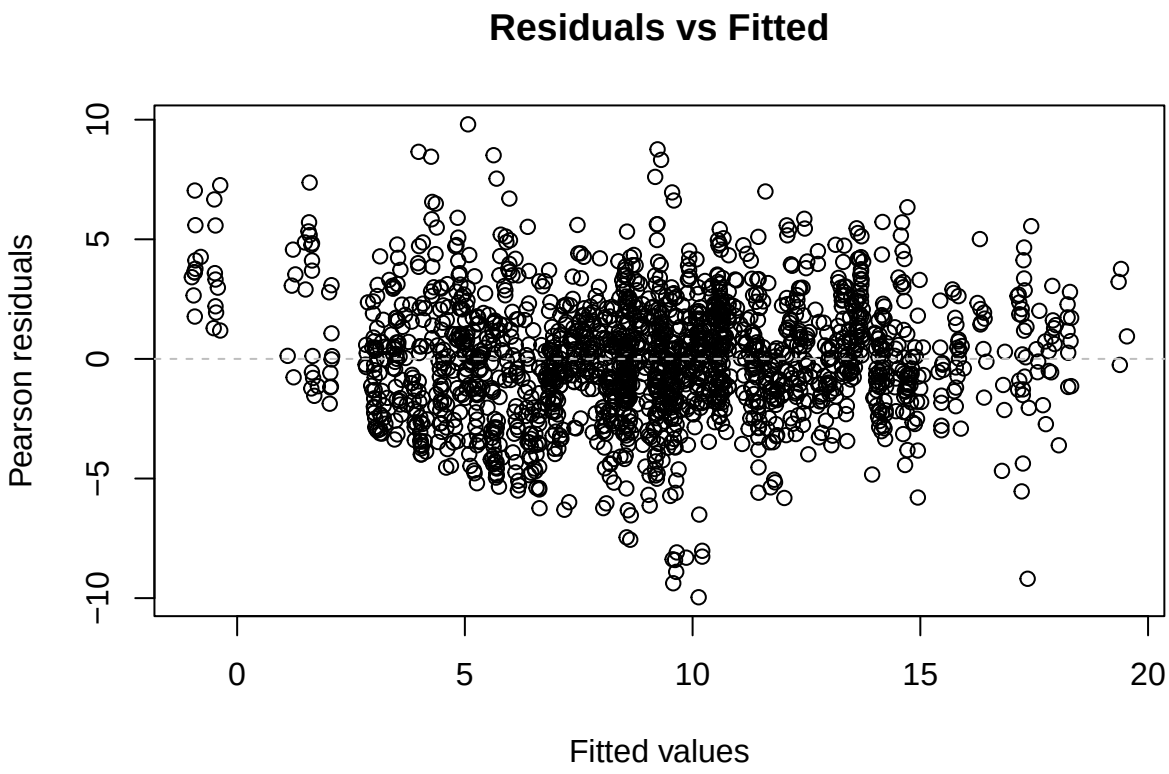
```
##
## Variables shrunk to 0 by Lasso (using lambda.min):

## None (at lambda.min)
```

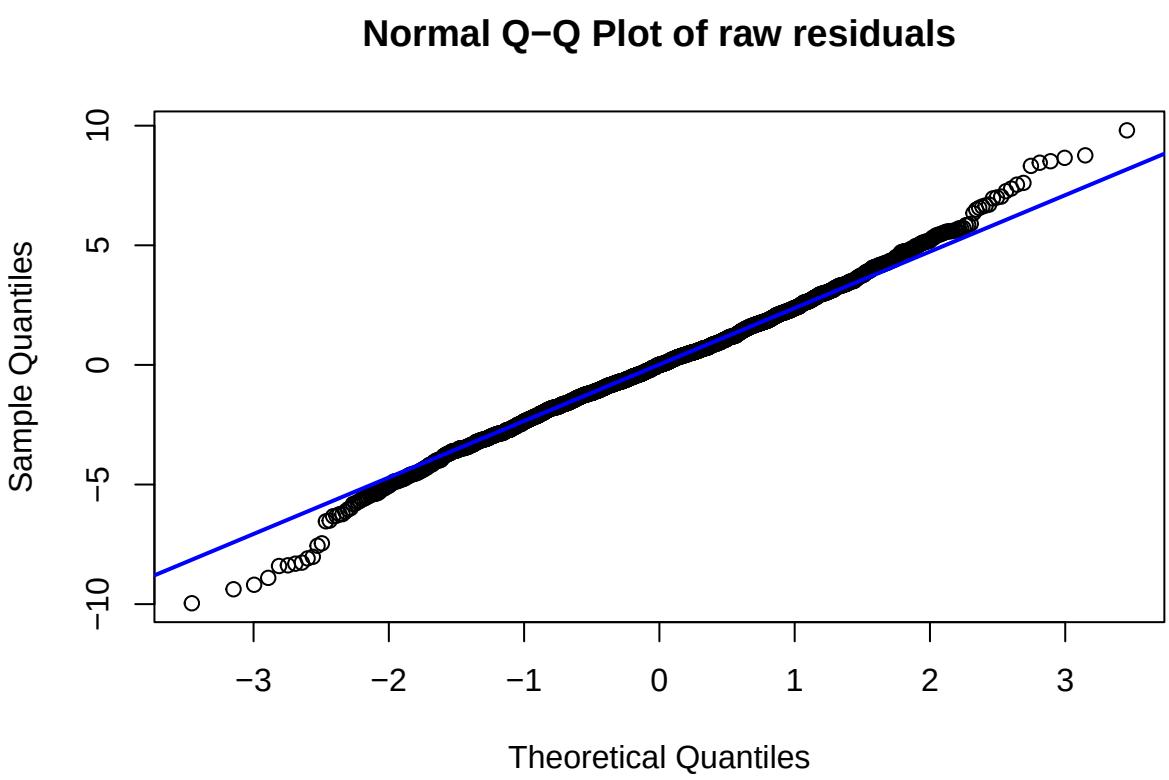
Based on the results above, the optimal λ value is 0.001, and all the variables in the final model are not shrunken to zero, meaning they are all considered important by Lasso.

Use Anova to Compare the Three Models

Model	AIC	BIC	Compared To	Chisq	Df	Pr(>Chisq)
m1	9318.7	9346.3	-	-	-	-
m2	8974.0	9040.1	m1	358.75	7	< 2.2e-16 ***
m4	9069.8	9119.4	m3	301.06	14	< 2.2e-16 ***
m3	8796.8	8923.5	m2	199.19	11	< 2.2e-16 ***

Based on the ANOVA results above, Model 3 has the lowest AIC and BIC among all the three models, and p-value of the chi-square test is less than 0.05, meaning adding those body measurements variables into the model 3 is statistically significant, also, by removing those non statistically significant variables in model 4, model 4 still perform poorer than model 3 based on the ANOVA results. Thus, model 3 is the overall best model. We can also plot the residual vs fitted plot and QQ plot to test for goodness-of-fit:


Residuals vs Fitted
Pearson residuals
10
5
0
−5
−10
0
5
10
15
20
Fitted values



Based on the two plots above, the residuals are randomly scattered around 0 and there is no pattern, which satisfies the heteroscedasticity assumption. Also, in qq plot, most of the residuals stay on the blue lines, indicating almost normal residuals. Both of the plots showed that the model fits well with the data.

Perform Two Sample T-test

Based on the two plots above, we can also consider using two sample t-test to show the difference in knee degrees change between the non-obese group and obese group for each walking stance is statistically significant.

```
## # A tibble: 6 x 11
##   Task  estimate estimate1 estimate2 statistic  p.value parameter conf.low
##   <fct>    <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>    <dbl>
## 1 FF       1.93      14.5      12.5      4.16  4.59e- 5      217.    1.01
## 2 OCHA     0.113      5.75      5.64     0.263 7.92e- 1      226.   -0.730
## 3 OCHB     3.18      11.7       8.55     7.89  6.58e-14      281.    2.39
## 4 OCMA     2.13       7.44      5.30     5.47  1.03e- 7      268.    1.37
## 5 OCMB     2.59      11.3       8.69     6.34  1.11e- 9      242.    1.79
## 6 PRF      2.32      12.0       9.69     5.20  4.79e- 7      212.    1.44
## # i 3 more variables: conf.high <dbl>, method <chr>, alternative <chr>
```

Based on the table above, the estimate shows the estimate for the difference between the two groups, estimate1 shows the mean knee angle changes for the normal group, estimate 2 shows the mean knee angle changes for the obese group, statistic shows the the t-statistic calculated, conf.low and conf.high show the confidence interval for the estimate, and p-value shows whether the t-test statistics are statistically significant

or not. Based on the table above, all the tasks showed statistically significant difference between normal and obese groups except OCHA. These results are consistent with the findings from the EDA plot.

Use Cross Validation to Test the Model Performance

We could also test the model's performance using 5-fold cross validation, where we split the training dataset into 5 folds and randomly pick one fold as validation set to test for the performance and use the other 4 folds to fit the model, we then take the average of the performance on each validation set at the end.

```
## $avg_RMSE
## [1] 2.59294
##
## $avg_R_squared
## [1] 0.6788659
##
## $fold_RMSEs
## [1] 2.621659 2.684303 2.631858 2.458529 2.568351
##
## $fold_R_squareds
## [1] 0.6620659 0.6813412 0.6751611 0.7100661 0.6656953
```

Based on the results above, the average root mean square error for linear mixed effect model is 2.59, and the average R-squared is 0.68. To validate and further improve our results, we can try GAMM(General Additive Mixed Models) and Bayesian Random Effects Models:

```
## $avg_RMSE
## [1] 2.593828
##
## $avg_R_squared
## [1] 0.6786574
##
## $fold_RMSEs
## [1] 2.613180 2.683352 2.633363 2.464417 2.574830
##
## $fold_R_squareds
## [1] 0.6642484 0.6815669 0.6747895 0.7086758 0.6640065
```

Based on the cross-validation results, GAMM gives very similar results with linear mixed effect models.

Based on the bayesian mixed effects model results, the average RMSE is 5.83, and the average R squared is -5.48, which suggest that it performs a lot worse than both GAMM and linear mixed effect models. Thus, based on all three models above, we plan to stick with the original linear mixed effects model.

Interpretation of the Model

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: InitialPeak ~ Group + age + Sex + Race + leg_l + DST + Stroop +
##     PA + Speed + neck_cir + SH_B + CH_B + HIP_B + HIP_D + ASIS +
##     L_thigh_cir + shank_cir + ankle_cir + ABSI + (1 | studyid) +
##     (1 | Task)
##    Data: df
```

```
##
## REML criterion at convergence: 8796.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.9044 -0.6182  0.0022  0.6298  3.8424
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  studyid  (Intercept) 24.101   4.909
##  Task     (Intercept)  7.351   2.711
##  Residual              6.511   2.552
## Number of obs: 1826, groups:  studyid, 25; Task, 6
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept) -11.48042   77.85110  -0.147
## Group         0.88102    1.86180   0.473
## age           0.12140    0.15572   0.780
## Sex          -0.44293    2.65443  -0.167
## Race          0.69062    0.56633   1.219
## leg_l         0.19914    0.10061   1.979
## DST           0.07188    0.17316   0.415
## Stroop        0.06365    0.60167   0.106
## PA            0.06258    0.73722   0.085
## Speed         1.30061    0.69608   1.868
## neck_cir      0.09850    0.50787   0.194
## SH_B          0.87595    0.17265   5.074
## CH_B         -0.54526    0.21456  -2.541
## HIP_B         0.31224    0.39539   0.790
## HIP_D         0.09220    0.35644   0.259
## ASIS          0.80981    0.43473   1.863
## L_thigh_cir  -0.44396    0.07986  -5.559
## shank_cir    -0.72106    0.29456  -2.448
## ankle_cir     0.41151    0.70122   0.587
## ABSI         -0.36599    0.10188  -3.592
```

Based on the model output above, the most important sections are **scaled residuals**, **random effects** and **fixed effects**. The **scaled residuals** showed that the residuals are symmetric around about 0 and satify the assumption of the linear mixed effect model. For the **random effects**, the results revealed significant variability attributable to differences between participants (Variance = 24.10, SD = 4.91) and, to a lesser extent, between tasks (Variance = 7.35, SD = 2.71). The residual variance, representing within-participant/task variability not explained by the model, was 6.51 (SD = 2.55).

For the **fixed effects**, the variables that showed statistically significant relationship with knee angle difference are SH_B(Shoulder Breadth), CH_B(Chest Breadth), L_thigh_cir(lower thigh circumference), shank_cir(shank circumference) and ABSI(a body shape index). We will interpret each of those coefficients below:

SH_B(Shoulder Breadth): Holding all other variables constant, one inch increase in shoulder breadth, the knee degrees change is expected to increase by 0.87.

CH_B(Chest Breadth): Holding all other variables constant, one inch increase in chest breadth, the knee degrees change is expected to decrease by 0.55.

L_thigh_cir(lower thigh circumference): Holding all other variables constant, one inch increase in

lower thigh circumference, the knee degrees change is expected to decrease by 0.44.

shank_cir(shank circumference): Holding all other variables constant, one inch increase in shank circumference, the knee degrees change is expected to decrease by 0.72.

ABSI(a body shape index): Holding all other variables constant, one inch increase in body shape index, the knee degrees change is expected to decrease by 0.37.

Conclusion

To summarize, the EDA part shows that non-obese people tend to bend their knees more when they walk compared to obese people, and this observation from plot is also supported the two sample t-tests results. The later linear mixed effect model further showed that Shoulder Breadth, Chest Breadth, lower thigh circumference, shank circumference and body shape index have statistically significant effects on the knee degrees change, among those variables, only shoulder breadth have positive relationship with the knee degrees change, and all the other variables are negatively correlated with the response variable.