Macalester Track Sprinters: 400 Meter Race Time Predictions

Hannah Sonsalla, Britney Millman, Kaitlyn Lindaman

Intro

Track and field is a sport which includes athletic competitions based on running, jumping, and throwing skills. It consists of a plethora of events with running races ranging from the 100 meter dash to the 10,000 meter run. Based on the events an athlete partakes in, they are considered either sprinters (short distance), distance runners, jumpers, hurdlers or vaulters. Our research project focuses on Macalester Track sprinters, the track athletes that compete in some of the shortest running events. We chose this topic as two of us are an avid part of the Macalester Track and Field program. We decided to look into 400 meter race times (sprinting event) among Macalester Track athletes as this distance tends to be popular and well-liked. Our goal was to create the most accurate model for predicting outdoor 400 meter race times while taking into consideration gender, where each athlete is from, indoor 400 meter time and indoor 200 meter time. From this, we formulated our research question: Do personal best 200 times and 400 times indoor correlate to personal best 400 times outdoor for Macalester track sprinters? Do best outdoor 400 times differ by gender or home region? We predicted that gender would be the most influential variable and the most reliable model would be based on indoor 200 meter and 400 meter race times and gender. Based on our results, we found the most influential variable to be indoor 400 meter race time, not gender. However, we found that the most accurate model for outdoor 400 meter race time included both indoor 400 meter race times and gender. Had our sample group been randomly selected, we would have been able to generalize our results to all Macalester track sprinters. However, since our sample was not randomly selected we were unable to generalize our results to the entire population.

Methods

We were able to find our data using observational studies (retrospective study using past statistics), since an experiment would be difficult to perform at this time. We collected the data for the race times from the website Apple Raceberry JaM, which keeps track of Minnesota Intercollegiate Athletic Conference track and field performances dating back to 1999. We went through indoor and outdoor performance lists through the year 2015 to identify and record the best indoor 200 meter, best indoor 400 meter, and best outdoor 400 meter race times for each applicable

Macalester athlete. We also made use of the Macalester Track and Field page on the Macalester Athletics website in order to find gender, graduation year, and home states of the athletes. After determining home states for each athlete, we then categorized each athlete by region based on their home states. Both of these resources, the Macalester Track and Field page on the Macalester Athletics website and the website Apple Raceberry JaM, are trustworthy as they are established institutions that have been recording this data for many years and have nothing to gain from misreporting this data.

Our variables for the data include: Name (athlete's name), Graduation Year (year athlete graduated or will graduate), gender (categorized as either male (M) or female (F)), home state (athlete's home state, or, if from abroad, home country), region (athlete's home region, categorized as Midwest (MW), West (W), Northeast (NE), Southwest (SW) or abroad (Abroad)), best indoor 200 meter race time, best indoor 400 meter race time and best outdoor 400 meter race time. All race times are recorded in seconds (ex. 26.27 seconds for an indoor 200 meter). Out of these, our categorical predictor variables are gender and region, and our quantitative predictor variables are best indoor 200 meter race time and best indoor 400 meter race time. Our response variable is best outdoor 400 meter race time.

The population of interest is all Macalester Track sprinters. Seeing as all sprinters do not run the 200 meter or 400 meter both indoor and outdoor, we chose a sample of all applicable athletes that we could find on the performance lists (those that had run at least a 200m indoor, 400m indoor and 400m outdoor during their athletic career at Macalester). We included athletes that have graduated and current athletes as we are looking for personal best times in the events. Our sample size consisted of 33 Macalester Sprinters, 19 men and 14 women. Even though our sample matches up with the desired population of Macalester track sprinters, we are unable to generalize the results to the population as our sample was not chosen randomly since we handpicked athletes.

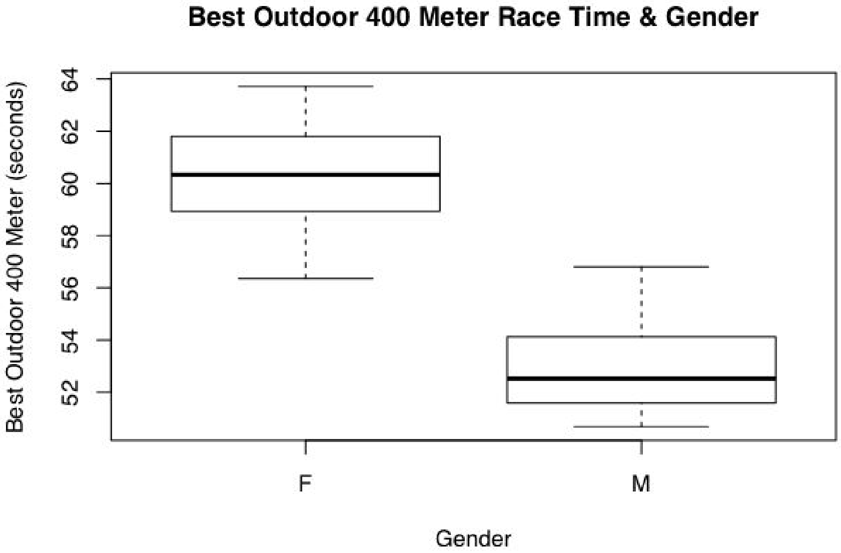
Since our sample was not chosen randomly, there are potential sources of sampling bias, namely, under-coverage and convenience sampling. Under-coverage bias occurs when individuals in the population have no chance of being in the sample population. Since the Apple Raceberry JaM website does not record performance lists before the year 2000 and the Macalester website does not provide roster access prior to 2000, we were unable to select any Macalester track sprinters that attended Macalester before the year 2000. These track sprinters had no chance of entering our sample. The convenience sampling bias occurs when a sample is chosen by convenience instead of by random chance. Our method of sampling did not include all of the track sprinters that we could potentially have found data for.

For this project we made use of RStudio (version 0.99.893), which is a free programming language for statistical computing. We made use of a variety of R functions: plot, boxplot, lm, summary and anova. The plot and boxplot functions are used to plot data according to specified variables. The lm command is used to fit linear models to the data. Summary serves as a way to display various information about the linear models such as coefficients, p-values, r-squared values etc. Lastly, we made use of the anova function during our hypothesis testing to determine significant differences between several models. The significance level we used was 0.05. This indicates that we are willing to accept up to a 5% risk of concluding that there is a relationship between certain variables when there actually is no such relationship. We chose this value as it is the norm and we felt that there was no substantial risk in mistakenly concluding that there is a relationship when there is not, nor was was there substantial risk in failing to recognize a significant relationship between our variables of interest.

Results

The variables we focused on were gender, indoor 200 meter race time, and indoor 400 meter race time. We did not utilize the home region of the athlete as we found that this was responsible for very little of the variability (~ 27 %) in 400 meter race times among athletes compared to the other variables. After taking other predictors into account, the home region of the athlete was never found to significantly increase the predictive ability of any model, so it is not included in any of the models examined in this paper.

First, we wanted to determine if there was a relationship between outdoor 400 meter race times and gender. We created side-by-side boxplots to see if there was a visual difference between males and females in respect to their 400 meter race times. There were two boxplots generated, one for female sprinters (F) and one for male sprinters (M).



The boxplot for female sprinters indicates that the median 400 meter race time is ~ 60.5 seconds whereas the median 400 meter race time for men is ~ 52.5 seconds. These values vary significantly. Additionally, it appears that the data for female sprinters is symmetric and unimodal as the median (thick black line) is spaced equally between the interquartile range (box). Q1 refers to the first quartile, or the first 25% of the data set. Q3 refers to the third quartile, or the first 75% of the data set. The interquartile range is the middle 50% of the data. For females, Q1 is ~ 59 and Q3 is ~ 62. This means that roughly half of Macalester women sprinters have a best outdoor 400 meter time between 59 and 62 seconds. On the other hand, the men's data is left skewed as the median is far closer to Q1 than Q3 and the IQR is located closer to the minimum non-outlier barrier. For males, Q1 is ~ 51.5 and Q3 is ~ 54.5. Roughly half of Macalester men sprinters have a best outdoor 400 meter time between 51.5 and 54.5 seconds. This plot is very relevant to our research question, as it displays that gender is a defining factor in outdoor 400 meter times.

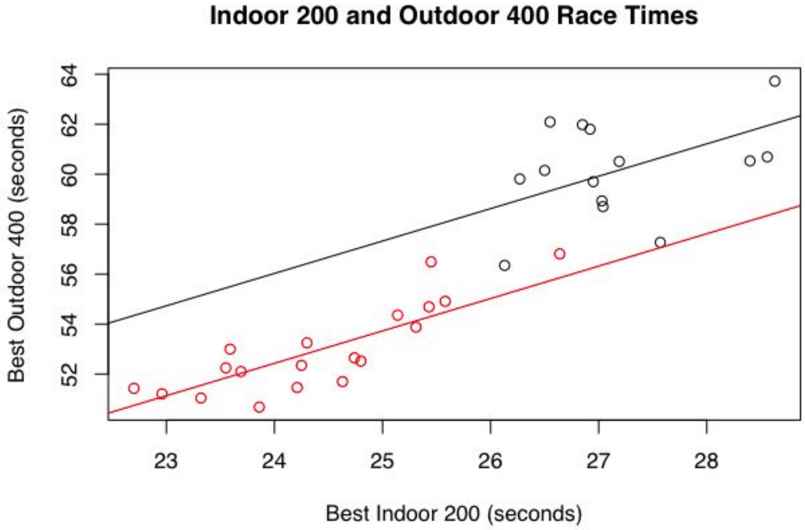
Then, we made a model, **groupWiseMeanGenderMod,** to find the numeric difference gender plays in race times. The regression equation for groupWiseMeanGenderMod is: **Predicted 400 meter race time = 60.159 + -7.171 I (Gender = M)**. It contains an indicator variable (I (Gender = M)). An indicator variable takes the value of 0 or 1 to indicate the presence or absence of a categorical effect that could potentially shift the outcome. In this case, the gender categorical variable takes a 1 if the athlete is male and 0 if the athlete is female to account for a gender difference in outdoor 400 meter times. The regression equation indicates that a women's outdoor

400 meter race time is predicted to be 60.159 seconds. If a man runs the outdoor 400 meter, his best time is predicted to be 7.171 seconds faster than women's (52.988 seconds). This value represents the difference in 400 meter race times due to gender. The r-squared value for this model is 0.7952. This means that 79.52% of the variation in the outdoor 400 meter times can be explained by gender difference.

We utilized hypothesis testing to determine whether or not using gender as a predictor was significantly more useful than simply using the overall mean to predict outdoor 400 meter times. The null hypothesis states that there is no difference in the mean outdoor 400 meter race time between males and females. The alternative hypothesis states that there is a difference in the mean outdoor 400 meter race time between males and females.

We assumed that the null hypothesis was true until we had enough evidence to indicate otherwise. We found such evidence by using an anova test. The anova test calculated our p-value to be 3.358*10^-12, which is less than our significance level of 0.05. This is enough evidence to reject the null hypothesis that mean outdoor 400 meter race times for males and females are equal in favor of the hypothesis that they are not equal. That is, the mean outdoor 400 meter race time for men is significantly different than the mean outdoor 400 meter race time for women. There is still a possibility that we are committing a Type 1 error, wherein we mistakenly reject the null hypothesis. There is a very small chance, 3.358*10^-12%, that we mistakenly rejected the null hypothesis.

Next, we plotted the best outdoor 400 times as predicted by the best indoor 200 times, by gender. In the below plot, the female athletes are indicated by black data points and the male athletes are indicated by red data points. This is a parallel lines model, where the slope of each line is the same, but the intercepts are different. This means that, as an athlete's best indoor 200 time increases by one second, their outdoor 400 time increases by the same amount regardless of gender. There is a difference of about 4 seconds between the two genders, indicated by their differing intercepts, ergo creating parallel lines. Since the data points are close to the generated linear models, we can see that the indoor 200 meter time is a good predictor of outdoor 400 meter times.



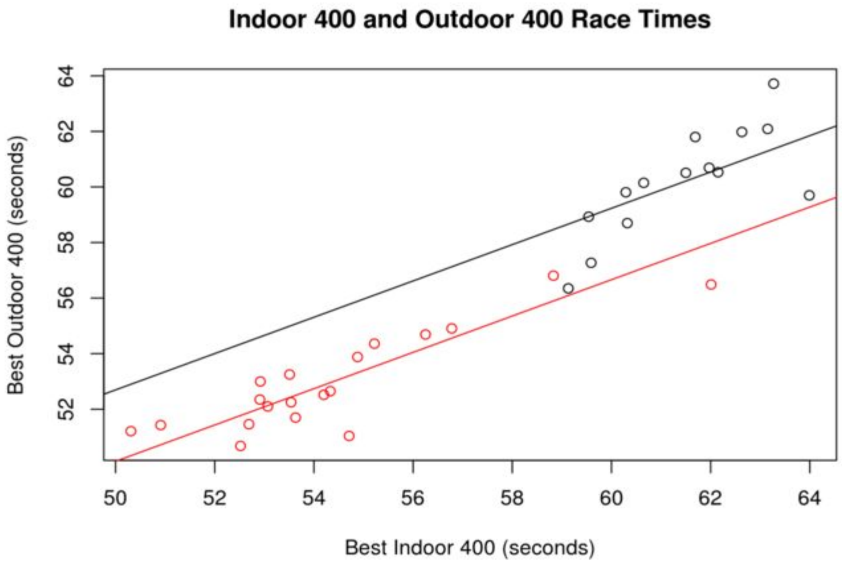
The linear regression line for the model, **gender200Mod**, is: **Predicted outdoor 400 time = 24.9107 + 1.2966(Best Indoor 200 time) + (-3.5973) * I(gender = M)**. This means that for every additional second on an athlete's indoor 200 meter time, their outdoor 400 meter time increases by 1.2966 seconds, regardless of gender. This model also predicts that men will run 3.5973 seconds faster than women. The r-squared value for this model is 0.8834, which means that 88.34% of the variation in outdoor 400 meter times can be explained by differences in indoor 200 meter times and gender.

We utilized hypothesis testing to determine whether or not using indoor 200 meter times as a predictor was significantly more useful than using only gender to predict outdoor 400 meter times. The null hypothesis states that, if gender is taken into account, the addition of 200 meter indoor times will not significantly improve the prediction of outdoor 400 meter times. The alternative hypothesis states that if gender is taken into account, the addition of 200 meter indoor times will significantly improve the prediction of outdoor 400 meter times.

We must assume that the null hypothesis is true until we have enough evidence to indicate otherwise. Once again, we found this evidence by using an anova test. The p-value is 4.553*10^-5, which is less than 0.05. This is enough evidence to reject the null hypothesis that if gender is taken into account, the addition of 200 meter indoor times will not significantly improve the prediction of outdoor 400 meter times, in favor of the alternative hypothesis. That is, if gender is taken into account, the addition of 200 meter indoor times will significantly improve the prediction of outdoor

400 meter times. There is a very small chance, 4.553*10^-5%, that we mistakenly rejected the null hypothesis.

Then, we plotted the relationship between indoor 400 meter, gender and outdoor 400 meter. This is another parallel lines model, in which linear models for both men and women have the same slope, but different intercepts. In the plot below, females athletes are indicated by black data points, and males are indicated by red data points. As shown, the intercept for females is approximately three seconds greater than the intercept for men. This means that on average females are predicted to run a slower time than men. Additionally, the slope for both linear models are positive, indicating that as best indoor 400 meter time decreases, best outdoor 400 meter time should decrease as well. Since the data points are close to the generated linear models, we can see that the indoor 400 meter time is a strong predictor of outdoor 400 meter time.



The linear regression line for this model, **gender400Mod** is: **Predicted Outdoor 400 Race Time = 20.01169 + 0.65366 * (Indoor 400 Race Time) + -2.56954 I (Gender = M)**. This model shows that for every additional second in an athlete's indoor 400 meter time, their outdoor 400 meter time will be 0.65366 seconds slower regardless of gender. This model also predicts that men will run 2.56954 seconds faster than women. The r-squared value for this model is 0.9279. This means that 92.79% of the variation in outdoor 400 meter times can be explained by the difference in indoor 400 meter times and gender.

Our hypothesis testing relating to this model was useful in determining whether or not including indoor 400 meter times as a predictor was significantly more useful than using only gender to predict outdoor 400 meter times. The null hypothesis for this test was that the addition of indoor 400 meter times will not significantly improve the prediction of outdoor 400 meter times, if gender is already accounted for. Alternatively, the alternative hypothesis was that the addition of indoor 400 meter times will significantly improve predictions of outdoor 400 meter times, assuming gender is already taken into account.

We assumed that the null hypothesis was correct until we had evidence to indicate otherwise. We made an anova test to find such evidence and found that the p-value for the test is 2.784*10^-08, which is less than 0.05. This is enough evidence to reject the null hypothesis if gender is taken into account, the addition of indoor 400 meter times will not significantly improve the prediction of outdoor 400 meter times, in favor of the alternative hypothesis. That is, if gender is taken into account, the addition of 400 meter indoor times will significantly improve the prediction of outdoor 400 meter times. There is still a possibility that we are committing a Type 1 error, wherein we mistakenly reject the null hypothesis. There is a very small chance, 2.784*10^-08%, that we mistakenly rejected the null hypothesis.

Finally, we created the model gender200400Mod using indoor 200 meter time, indoor 400 meter time, and gender as non-interactive predictors for outdoor 400 meter time. The linear regression line for **gender200400Mod** is: **Predicted Outdoor 400 Race Time = 16.0175 + 0.5421 (Indoor 400 Race Time) + 0.3989 (Indoor 200 Race Time) - 2.2552 I (Gender = M)**. This means that as an athlete's indoor 400 race time increases by one second, if their indoor 200 race time does not change, their outdoor 400 race time is predicted to increase by 0.5421 seconds. If an athlete's indoor 200 race time increases by one second and their indoor 400 race time does not change, then their outdoor 400 race time is predicted to increase by 0.3989 seconds. This model predicts that men will run 2.2552 seconds faster than women who have run the same indoor 200 and indoor 400 races, thus demonstrating the difference in intercepts between the parallel lines. The r-squared value for this model is 0.9324. This means that 93.24% of the variation in outdoor 400 meter times can be explained by the difference in indoor 400 meter times, indoor 200 meter times and gender.

We utilized hypothesis testing to determine whether or not using gender and both indoor 200 meter times and indoor 400 meter times as predictors was significantly more useful than using gender and indoor 200 meter times or gender and indoor 400 meter times to predict outdoor 400 meter times. Our first hypothesis test was to compare **gender200Mod** (gender and indoor 200 meter time as predictors) against **gender200400Mod** (gender, indoor 200 meter times and indoor 400 meter times). The null hypothesis for this test stated that, if gender and indoor 200 times were accounted for, the addition of indoor 400 meter times will not significantly improve the prediction of outdoor 400 meter times. The alternative hypothesis states that if gender and indoor 200 times were accounted for, the addition of indoor 400 meter times will significantly improve the prediction of outdoor 400 meter times.

We assumed that the null hypothesis in this case was correct until we had evidence to indicate otherwise. An anova test was used to find such evidence and found that the p-value for the test is 8.008*10^-05, which is less than 0.05. This is enough evidence to reject the null hypothesis in favor of the alternative hypothesis. That is, if gender and indoor 200 meter time is taken into account, the addition of indoor 400 meter times will significantly improve the prediction of outdoor 400 meter times. There is still a possibility that we are committing a Type 1 error, wherein we mistakenly reject the null hypothesis. There is a very small chance, 8.008*10^-05%, that we mistakenly rejected the null hypothesis.

On the other hand, our hypothesis testing revealed that including the indoor 200 meter time to **gender400Mod** which already includes the indoor 400 meter time and gender, does not significantly improves the predicted outdoor 400 meter times. The null hypothesis for this test was that the addition of indoor 200 meter times will not significantly improve the prediction of outdoor 400 meter times, if the 400 meter times and gender are already accounted for. Alternatively, the alternative hypothesis was that the addition of indoor 200 meter times will significantly improve predictions of outdoor 400 meter times, assuming gender and indoor 200 meter times are already taken into account.

We assumed that the null hypothesis was correct until we had evidence to indicate otherwise. We made an anova test to find such evidence and found that the p-value for this test is 0.1761, which is more than 0.05. This is not enough evidence to reject the null hypothesis that, if gender and the indoor 400 meter times are taken into account, the addition of indoor 200 meter times will not significantly improve the prediction of outdoor 400 meter times. That is, if gender and the indoor 400 meter times are taken into account, the addition of 200 meter indoor times will not significantly improve the prediction of outdoor 400 meter times. Since **gender200400Mod** is more complex, but does not significantly increase our predictive ability of outdoor 400 meter times, we will reject it in favor of our less complex model, **gender400Mod** that includes the indoor 400 meter times and gender. There is the possibility, however, that we have mistakenly failed to reject the null hypothesis, a Type 2 error, but this possibility is small.

Discussion

We found that the most significant predictors of outdoor 400 meter times were indoor 400 meter times and gender. In other words, if we would like to predict a Macalester track sprinters outdoor 400 meter time, we should use their indoor 400 meter time and gender to create the most accurate representation. We could be committing a Type 2 error, where one mistakenly fails to reject the null hypothesis and therefore assume that the less complex model, the one including only the indoor 400 meter times and gender as predictors, is the best model for predicting outdoor 400 meter times, instead of a more complex model that also takes into account indoor 200 meter times. Our sample size, only 33 cases, is relatively small, which could increase the chances of mistakenly failing to reject the null hypothesis, a Type 1 error. Our research may be limited as there have not been many track sprinters in the Macalester track and field program. Since our sample was not randomly selected we are unable to generalize our results to the entire population of Macalester track sprinters. For future research, we would like to see if the number of hours an athlete sleeps affects outdoor 400 meter performance and if the type of activity an athlete partakes in on active recovery days (rest days that include activities that are less intense, and have less volume) effects this as well. It would be interesting to investigate these two potential predictors because they are within the athlete's control to change. If we find a significant relationship between these possible predictors and outdoor 400 meter race times, athletes can use this information to possibly better their race times.

Appendix

Grand Mean Model **(grandMeanMod)***:*

```
grandMeanMod < - (Best_Outdoor_400~1)
summary(grandMeanMod)
```

```
##
## Call:
## lm(formula = Best_Outdoor_400 ~ 1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.351 -3.681 -1.121  3.779  7.689
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  56.0306     0.7026   79.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.036 on 32 degrees of freedom
```

The above r code was used to determine mean outdoor 400 race time.

*Groupwise Mean Model by Region (***groupWiseMeanRegionMod)***:*

```
groupWiseMeanRegionMod <- lm(Best_Outdoor_400~Region)
summary(groupWiseMeanRegionMod)
```

```
##
## Call:
## lm(formula = Best_Outdoor_400 ~ Region)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.060 -2.513 -0.170  2.536  7.878
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  54.2125     1.3062  41.505   <2e-16 ***
## RegionMW      0.6475     1.6862   0.384   0.7039
## RegionNE      3.2775     2.1061   1.556   0.1309
## RegionSW      9.5075     3.9185   2.426   0.0219 *
## RegionW       3.7618     1.9120   1.967   0.0591 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.694 on 28 degrees of freedom
## Multiple R-squared:  0.2669, Adjusted R-squared:  0.1621
## F-statistic: 2.548 on 4 and 28 DF,  p-value: 0.06137
```

The above r code was used to determine the r-squared value for the groupwise mean model by region.

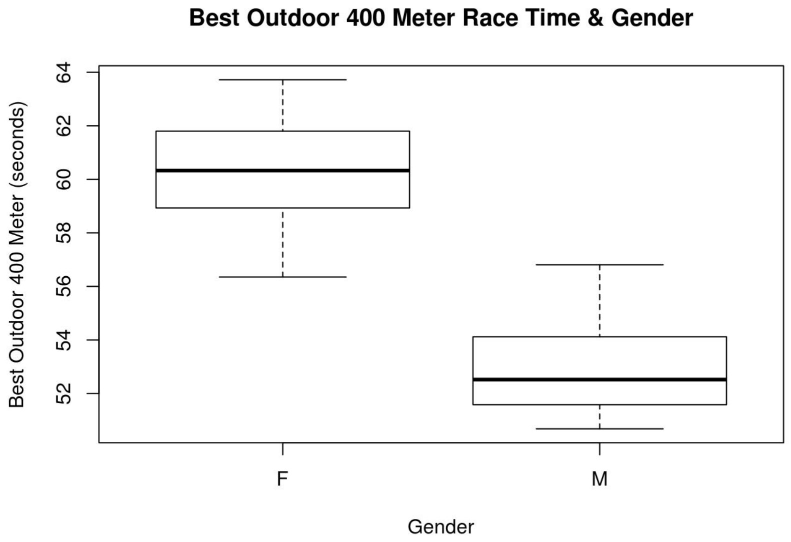
```
anova(grandMeanMod, groupWiseMeanRegionMod)
```

```
## Analysis of Variance Table
##
## Model 1: Best_Outdoor_400 ~ 1
## Model 2: Best_Outdoor_400 ~ Region
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     32 521.26
## 2     28 382.15  4    139.11 2.5481 0.06137 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The above r code was used to generate the p-value to determine the significance of the groupwise mean model by region.

*Groupwise Mean Model by Gender (***groupWiseMeanGenderMod***):*

```
boxplot(Best_Outdoor_400 ~ Gender, ylab = "Best Outdoor 400 Meter (seconds)", xlab = "Gender", main = "Best Outdoor 400 Meter Race Time & Gender")
```



The above r code was used to generate the side-by-side boxplot “Best Outdoor 400 Meter Race Time & Gender”.

```
groupWiseMeanGenderMod<-lm(Best_Outdoor_400~Gender)
summary(groupWiseMeanGenderMod)
##
## Call:
## lm(formula = Best_Outdoor_400 ~ Gender)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8093 -1.2884 -0.3384  1.3716  3.8216
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  60.1593     0.4960  121.29  < 2e-16 ***
## GenderM      -7.1709     0.6537  -10.97 3.36e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.856 on 31 degrees of freedom
## Multiple R-squared:  0.7952, Adjusted R-squared:  0.7886
## F-statistic: 120.3 on 1 and 31 DF,  p-value: 3.358e-12
```

The above r code was used to determine the regression equation and r-squared value for the groupwise mean model by gender.

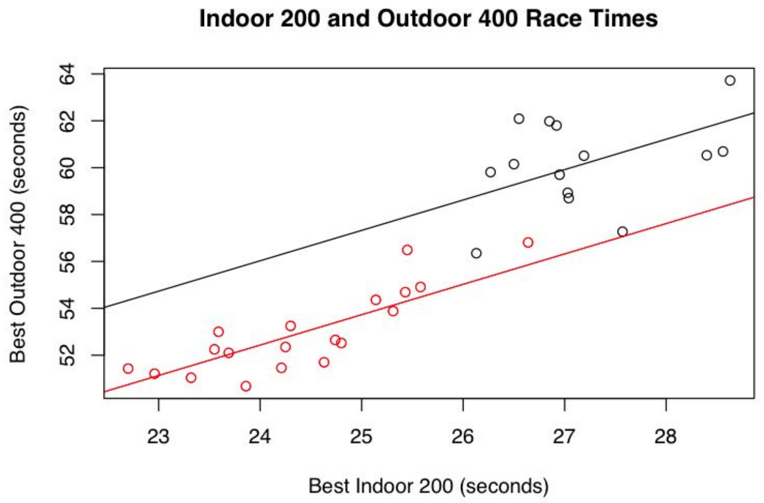
```
anova(grandMeanMod, groupWiseMeanGenderMod)
```

```
## Analysis of Variance Table
##
## Model 1: Best_Outdoor_400 ~ 1
## Model 2: Best_Outdoor_400 ~ Gender
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     32 521.26
## 2     31 106.77  1    414.49 120.34 3.358e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The above r code was used to generate the p-value to determine the significance of the groupwise mean model by gender.

Parallel Lines Model by Indoor 200 Meter Race Time and Gender **(gender200Mod)***:*

```
plot(Best_Outdoor_400~Best_Indoor_200, xlab = "Best Indoor 200 (seconds)", ylab = "Best Outdoor 400 (seconds)", main = "Indoor 200 and Outdoor 400 Race Times", col = Gender)
abline(24.9107, 1.2966)
abline(24.9107-3.5973, 1.2966, col = "red")
```



The above r code was used to generate the parallel lines plot “Indoor 200 and Outdoor 400 Race Times”.

```
gender200Mod <-lm(Best_Outdoor_400~Best_Indoor_200+Gender)
summary(gender200Mod)
##
## Call:
## lm(formula = Best_Outdoor_400 ~ Best_Indoor_200 + Gender)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3885 -1.0283  0.1262  0.8371  2.7541
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      24.9107     7.4112   3.361 0.002130 **
## Best_Indoor_200   1.2966     0.2723   4.762 4.55e-05 ***
## GenderM          -3.5973     0.9025  -3.986 0.000397 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Residual standard error: 1.424 on 30 degrees of freedom
## Multiple R-squared:  0.8834, Adjusted R-squared:  0.8756
## F-statistic: 113.6 on 2 and 30 DF,  p-value: 1.007e-14
```

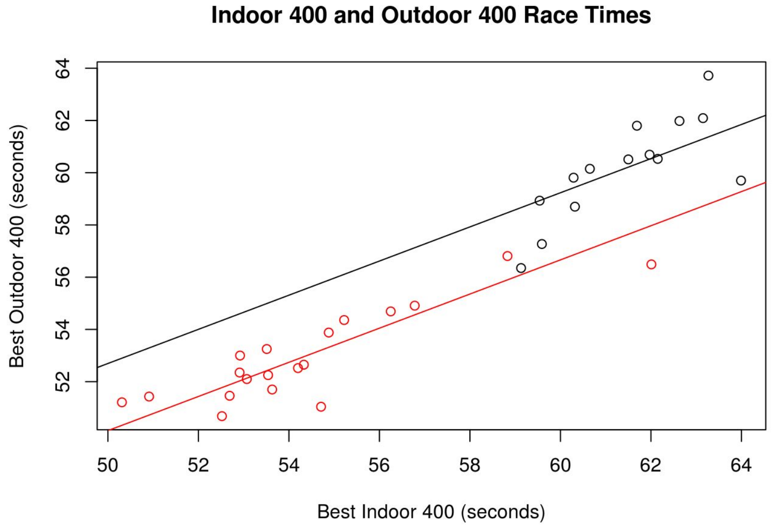
The above r code was used to generate the regression equation and r-squared value for the parallel lines model by indoor 200 meter race time and gender.

```
anova(groupWiseMeanGenderMod, gender200Mod)
## Analysis of Variance Table
##
## Model 1: Best_Outdoor_400 ~ Gender
## Model 2: Best_Outdoor_400 ~ Best_Indoor_200 + Gender
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     31 106.775
## 2     30  60.805  1     45.97 22.681 4.553e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The above r code was used to generate the p-value to determine the significance of the parallel lines model by indoor 200 meter race time and gender.

Parallel Lines Model by Indoor 400 Meter Race Time and Gender **(gender400Mod)***:*

```
plot(Best_Outdoor_400~Best_Indoor_200, xlab = "Best Indoor 200 (seconds)", ylab = "Best Outdoor 400 (seconds)", main = "Indoor 200 and Outdoor 400 Race Times", col = Gender)
abline(24.9107,1.2966)
abline(24.9107-3.5973,1.2966, col = "red")
```



The above r code was used to generate the parallel lines plot “Indoor 400 and Outdoor 400 Race Times”.

```
gender400Mod<-lm(Best_Outdoor_400~Best_Indoor_400+Gender)
summary(gender400Mod)
##
## Call:
## lm(formula = Best_Outdoor_400 ~ Best_Indoor_400 + Gender)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3129 -0.4237  0.2979  0.7994  2.3510
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     20.01169    5.41039   3.699 0.000868 ***
## Best_Indoor_400  0.65366    0.08795   7.432 2.78e-08 ***
## GenderM         -2.56954    0.73400  -3.501 0.001474 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.119 on 30 degrees of freedom
## Multiple R-squared:  0.9279, Adjusted R-squared:  0.9231
## F-statistic:   193 on 2 and 30 DF,  p-value: < 2.2e-16
```

The above r code was used to generate the regression equation and r-squared value for the parallel lines model by indoor 400 meter race time and gender.

```
anova(groupWiseMeanGenderMod, gender400Mod)
```

```
## Analysis of Variance Table
##
## Model 1: Best_Outdoor_400 ~ Gender
## Model 2: Best_Outdoor_400 ~ Best_Indoor_400 + Gender
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     31 106.775
## 2     30  37.583  1    69.192 55.232 2.784e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The above r code was used to generate the p-value to determine the significance of the parallel lines model by indoor 400 meter race time and gender.

Model with Indoor 200 Meter Race Time, Indoor 400 Meter Race Time, and Gender **(gender200400Mod)***:*

```
gender200400Mod<-lm(Best_Outdoor_400~Best_Indoor_400+Best_Indoor_200+Gender)
summary(gender200400Mod)
```

```
##
## Call:
## lm(formula = Best_Outdoor_400 ~ Best_Indoor_400 + Best_Indoor_200 +
##     Gender)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1473 -0.5249  0.1613  0.6810  1.9809
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      16.0175     6.0578   2.644  0.01308 *
## Best_Indoor_400   0.5421     0.1182   4.586 8.01e-05 ***
## Best_Indoor_200   0.3989     0.2877   1.387  0.17613
## GenderM          -2.2552     0.7577  -2.977  0.00583 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.102 on 29 degrees of freedom
## Multiple R-squared:  0.9324, Adjusted R-squared:  0.9254
## F-statistic: 133.3 on 3 and 29 DF,  p-value: < 2.2e-16
```

The above r code was used to generate the regression equation and r-squared value for the parallel lines model by indoor 200 meter race time, 400 meter race time, and gender.

```
anova(gender200Mod, gender200400Mod)
## Analysis of Variance Table
##
## Model 1: Best_Outdoor_400 ~ Best_Indoor_200 + Gender
## Model 2: Best_Outdoor_400 ~ Best_Indoor_400 + Best_Indoor_200 + Gender
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1     30 60.805
## 2     29 35.246  1    25.559 21.03 8.008e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The above r code was used to generate the p-value to determine the significance of the parallel lines model (based on 200 meter indoor times) by indoor 200 meter race time, indoor 400 meter race time, and gender.

```
anova(gender400Mod,gender200400Mod)

## Analysis of Variance Table
##
## Model 1: Best_Outdoor_400 ~ Best_Indoor_400 + Gender
## Model 2: Best_Outdoor_400 ~ Best_Indoor_400 + Best_Indoor_200 + Gender
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     30 37.583
## 2     29 35.246  1    2.3367 1.9227 0.1761
```

The above r code was used to generate the p-value to determine the significance of the parallel lines model (based on 400 meter indoor times) by indoor 200 meter race time, indoor 400 meter race time, and gender.