Code Sample (ED Project)

Annika Kufrovich

Library Loading

```
library(dplyr)
library(nlme)
library(ggplot2)
library(car)
```

Intro

For this project, I wanted to model disordered eating beliefs. This uses data from a study of college students over the course of their first semester, looking at disordered eating beliefs, among other variables like gender, athlete status, a contingencies of self-worth (CSW) scale, and hours exercised per week. The surveys were taken before, during, and at the end of the semester, so I will be looking at what can be used to predict disordered eating beliefs and if there is change over time.

Cleaning the Data

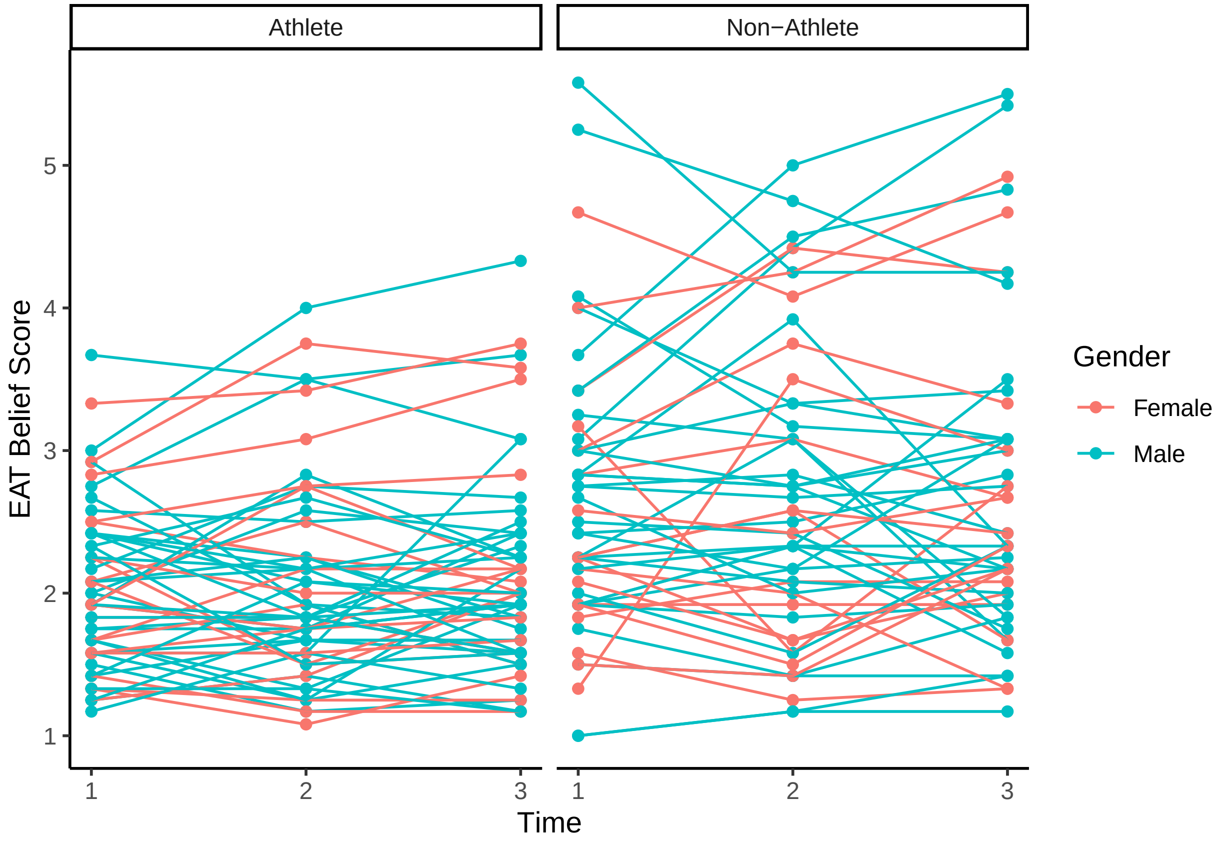
There is a decent amount of missing data. Some participants did not take the 2nd and/or 3rd survey. I removed these incomplete cases for a clear model over time. However, in the future, I would consider different means of addressing these missing data points. We can see that we have 66 people who are missing the 3rd survey and 57 people missing the 2nd survey based on the time sums. There is no one missing both surveys.

```
#Recoding categorical variables appropriately
ed <- read.csv("ED.csv")
ed.clean <- ed %>%
  mutate(Gender = as.factor(ifelse(Gender == 1, "Male", "Female"))) %>%
  mutate(Athlete = as.factor(ifelse(Athlete == 1, "Athlete", "Non-Athlete")))


##identifying incomplete cases by summing time
ed.incomp <- ed.clean %>%
  mutate(Time2 = ifelse(is.na(EATbelief) == TRUE, NA, Time))%>%
  group_by(ID) %>%
  summarize(timesum = sum(Time2, na.rm = TRUE))

##Missing counts (6 is a complete case)
missing.table <-summary(as.factor(ed.incomp$timesum))

##removing incomplete cases
ed.cleanjoin <- left_join(ed.clean, ed.incomp, by = "ID") %>%
  filter(timesum == 6)

##Removing extraneous dependent variables from dataset
ed.clean2 <- ed.cleanjoin %>%
  select(ID, Time, Gender, Athlete, CSWapp, EATbelief, ExerciseHrsWeek)

##change over time may not be constant, making a categorical time variable
ed.clean2$Timefact <- as.factor(ed.clean2$Time)
```

Plotting The data

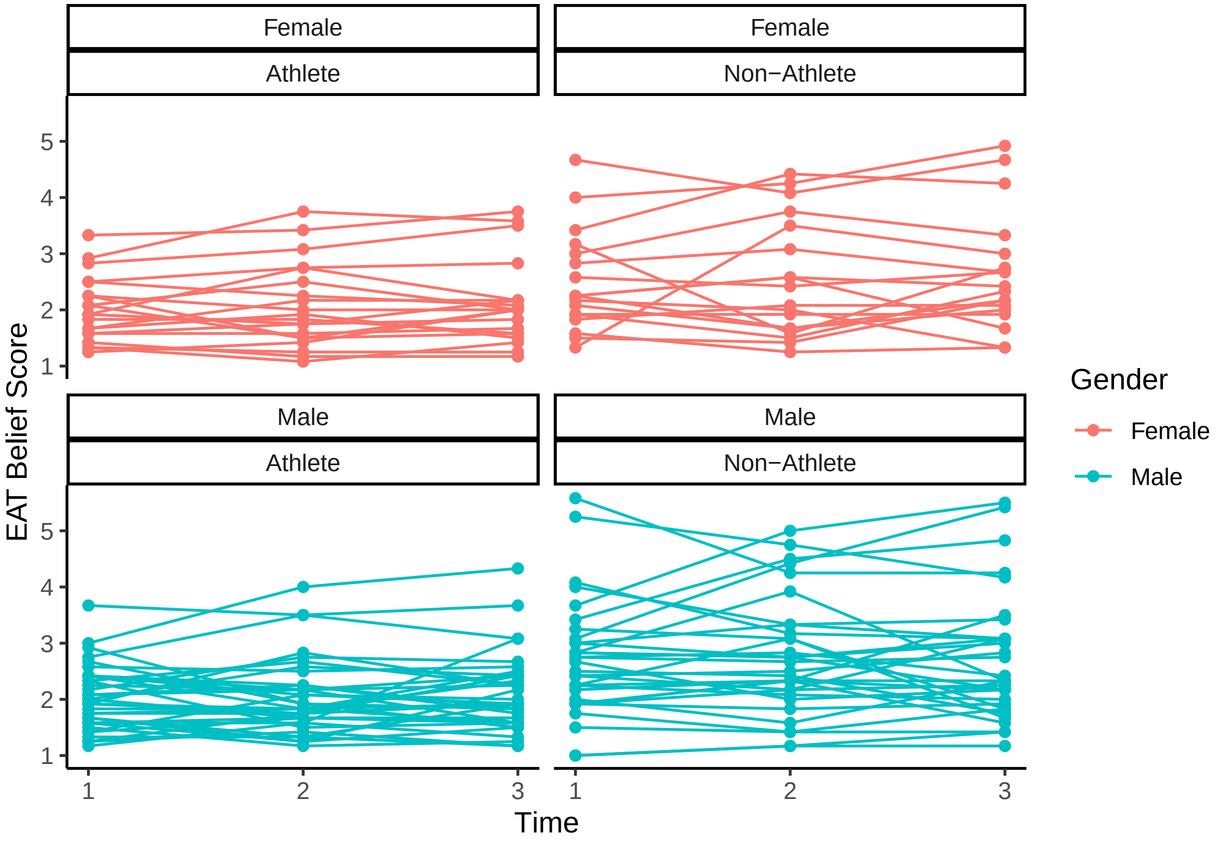
Here's my initial plot, even splitting by Athlete status this is a bit cluttered and confusing, so lets try splitting the plots by gender as well.

```
Spaghetti.Plot <- ggplot(ed.clean2, aes(x = Time,
                                       y = EATbelief,
                                       color = Gender,
                                       group = ID)) +
  geom_point() +
  geom_line() +
  facet_wrap(vars(Athlete)) +
  theme(legend.position = "top") +
  theme_classic() +
  labs(y = "EAT Belief Score") +
  scale_x_continuous(breaks = seq(1:3))
Spaghetti.Plot
```



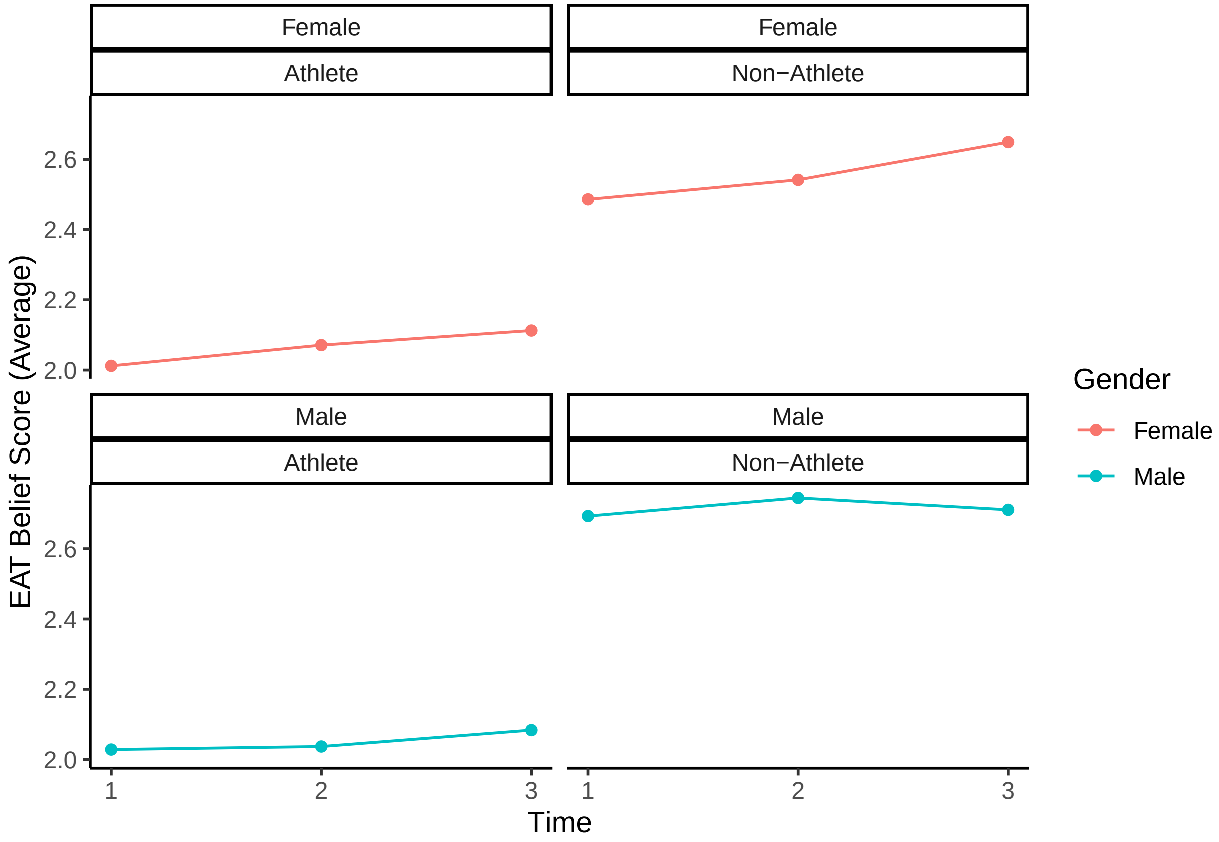
Though clearer, this next plot is still fairly cluttered. Let's look at averages for each of the subdivisions in the following plot.

```
Spaghetti.Plot1 <- ggplot(ed.clean2, aes(x = Time,
                                         y = EATbelief,
                                         color = Gender,
                                         group = ID)) +
  geom_point() +
  geom_line() +
  facet_wrap(vars(Gender, Athlete)) +
  theme(legend.position = "top") +
  theme_classic() +
  labs(y = "EAT Belief Score") +
  scale_x_continuous(breaks = 1:3)
Spaghetti.Plot1
```



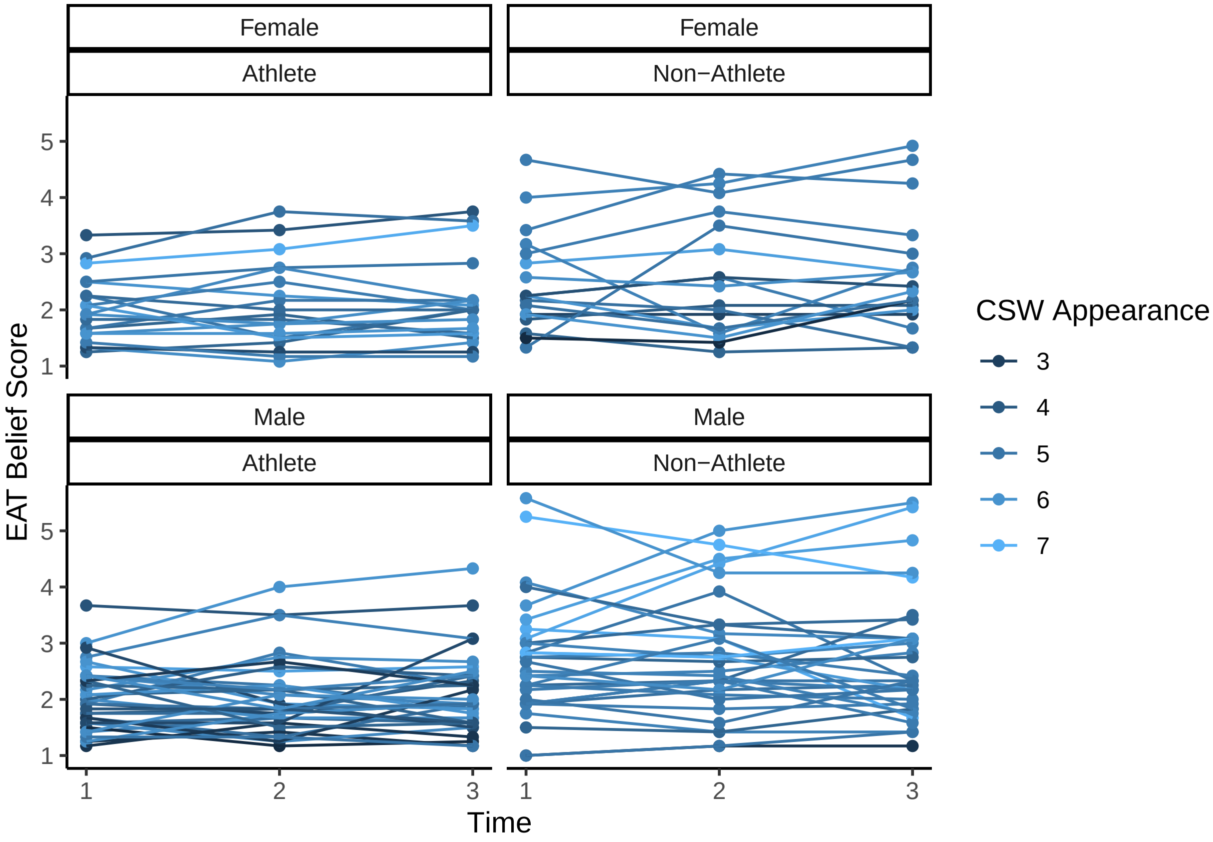
This is a plot of the means in each group over time. It seems there is likely a difference between athletes and non-athletes and potentially a small interaction between gender and athlete status, but it is difficult to be totally sure from these visuals.

```
##summarizing the data for plotting
ed.cleansum <- ed.clean2 %>%
  group_by(Gender, Athlete, Time) %>%
  summarise(EATbelief.mean = mean(EATbelief, na.rm = TRUE))
ed.cleansum$ID <- rep(1:4, each = 3)

Spaghetti.Plot2 <- ggplot(ed.cleansum, aes(x = Time,
                                          y = EATbelief.mean,
                                          color = Gender,
                                          group = ID)) +
  geom_point() +
  geom_line() +
  facet_wrap(vars(Gender, Athlete)) +
  theme(legend.position = "top") +
  theme_classic() +
  labs(y = "EAT Belief Score (Average)") +
  scale_x_continuous(breaks = 1:3)
Spaghetti.Plot2
```



Quick check for any interactive patterns with CSWapp. Hard to tell with these plots

```
Spaghetti.Plot3 <- ggplot(ed.clean2, aes(x = Time,
                                         y = EATbelief,
                                         color = CSWapp,
                                         group = ID)) +
  geom_point() +
  geom_line() +
  facet_wrap(vars(Gender, Athlete)) +
  theme_classic() +
  labs(y = "EAT Belief Score") +
  scale_x_continuous(breaks = 1:3) +
  guides(color = guide_legend(title = "CSW Appearance"))
Spaghetti.Plot3
```



Modeling EATbelief

Since this data looks at a variety of subjects over time, I will be fitting a linear mixed model, specifically a random intercepts model. This allows us to account for randomness of each participant's starting point while allowing us to fit a linear model for the general effect of each variable.

I started with a large model including several interaction terms and eventually reduced it through Anova and Likelihood ratio tests shown below, with `ri.ed3cor` being the final model.

```
ri.ed <- lme(EATbelief ~ Timefact + Gender + Athlete +
               CSWapp + ExerciseHrsWeek +
               Timefact*Athlete + Timefact*Gender + Gender*Athlete +
               Timefact*ExerciseHrsWeek,
             random = ~1|ID,
             data = ed.clean2)

ri.eds <- lme(EATbelief ~ Timefact + Athlete +
                CSWapp + ExerciseHrsWeek + Gender,
             random = ~1|ID,
             na.action = na.omit,
             data = ed.clean2)



c(summary(ri.ed)$BIC, summary(ri.ed)$AIC)
```

```
## [1] 708.7106 649.2382
```

```
c(summary(ri.eds)$BIC, summary(ri.eds)$AIC)
```

```
## [1] 646.0643 612.4062
```

```
LRT <- 2*(ri.eds$logLik - ri.ed$logLik)
1 - pchisq(LRT, df = 7)
```

```
## [1] 0.001823321
```

```
Anova(ri.ed)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: EATbelief
##                             Chisq Df Pr(>Chisq)
## Timefact                   1.5464  2  0.4615257
## Gender                     0.2954  1  0.5867528
## Athlete                   11.7343  1  0.0006136 ***
## CSWapp                    14.5046  1  0.0001398 ***
## ExerciseHrsWeek            0.1814  1  0.6701739
## Timefact:Athlete           0.7676  2  0.6812545
## Timefact:Gender            0.6296  2  0.7299168
## Gender:Athlete             0.1029  1  0.7484218
## Timefact:ExerciseHrsWeek   1.2545  2  0.5340475
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

From our tests it seems none of the above interactive effects contribute meaningfully to our model individually. Overall, the model is a better fit based on the Likelihood ratio test but this could be incidental. It also seems like there may not be a substantial effect of Time, Gender, or Hours exercised per week, but we’ll get back to those. Next, we’ll see if any interactions with CSWapp add anything to our model.

```
ri.ed2 <- lme(EATbelief ~ Timefact + Athlete +
                CSWapp + ExerciseHrsWeek + Gender +
                CSWapp*Timefact + CSWapp*Gender + CSWapp*Athlete,
             random = ~1|ID,
             na.action = na.omit,
             data = ed.clean2)

c(summary(ri.ed2)$BIC, summary(ri.ed2)$AIC)
```

```
## [1] 671.0784 622.6294
```

```
c(summary(ri.eds)$BIC, summary(ri.eds)$AIC)
```

```
## [1] 646.0643 612.4062
```

```
LRT2 <- 2*(ri.eds$logLik - ri.ed2$logLik)
1 - pchisq(LRT2, df = 4)
```

```
## [1] 0.694786
```

```
Anova(ri.ed2)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: EATbelief
##                    Chisq Df Pr(>Chisq)
## Timefact          1.5674  2   0.456719
## Athlete          10.4866  1   0.001202 **
## CSWapp           16.1609  1  5.818e-05 ***
## ExerciseHrsWeek   0.1134  1   0.736283
## Gender            0.0082  1   0.927914
## Timefact:CSWapp   0.3987  2   0.819273
## CSWapp:Gender     2.3777  1   0.123077
## Athlete:CSWapp    9.1815  1   0.002445 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

It looks like all of these interactions do not significantly add to our model. Even though the interaction of Athlete status and CSWapp appears significant individually, its overall contribution to the model is not significant enough to justify keeping it. Now let’s look at dropping gender and Hours exercised per week from our model.

```
ri.ed3 <- lme(EATbelief ~ Timefact + Athlete + CSWapp,
             random = ~1|ID,
             na.action = na.omit,
             data = ed.clean2)

c(summary(ri.ed3)$BIC, summary(ri.ed3)$AIC)
```

```
## [1] 624.9795 598.7561
```

```
c(summary(ri.eds)$BIC, summary(ri.eds)$AIC)
```

```
## [1] 646.0643 612.4062
```

```
LRT3 <- 2*(ri.ed3$logLik - ri.eds$logLik)
1 - pchisq(LRT3, df = 2)
```

```
## [1] 0.008025924
```

```
Anova(ri.eds)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: EATbelief
##                    Chisq Df Pr(>Chisq)
## Timefact          1.5792  2  0.4540308
## Athlete          11.5363  1  0.0006825 ***
## CSWapp           14.9937  1  0.0001079 ***
## ExerciseHrsWeek   0.1187  1  0.7304279
## Gender            0.3016  1  0.5829030
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Given this is the 3rd time using the likelihood ratio, we are more likely to get a false positive so I will not be considering that value as terribly strong $\alpha_{new} = .05^3 = .00125$. Because the AIC and BIC are better for the smaller model and the coefficients for gender and Hours Exercised are not significant on their own I think its best to leave them out of the model. For clarity here is what the model now looks like:

```
ri.ed3 <- lme(EATbelief ~ Timefact + Athlete + CSWapp,
              random = ~1|ID,
              data = ed.clean2)


summary(ri.ed3)
```

```
## Linear mixed-effects model fit by REML
##   Data: ed.clean2
##        AIC      BIC   logLik
##   598.7561 624.9795 -292.378
##
## Random effects:
##  Formula: ~1 | ID
##         (Intercept)  Residual
## StdDev:   0.6994372 0.4064073
##
## Fixed effects:  EATbelief ~ Timefact + Athlete + CSWapp
##                         Value Std.Error  DF  t-value p-value
## (Intercept)         0.7110428 0.3542284 210 2.007300  0.0460
## Timefact2           0.0390566 0.0558243 210 0.699634  0.4849
## Timefact3           0.0707547 0.0558243 210 1.267453  0.2064
## AthleteNon-Athlete 0.5092579 0.1456438 103 3.496599  0.0007
## CSWapp              0.2707906 0.0701457 103 3.860405  0.0002
##  Correlation:
##                     (Intr) Tmfct2 Tmfct3 AthN-A
## Timefact2           -0.079
## Timefact3           -0.079  0.500
## AthleteNon-Athlete -0.027  0.000  0.000
## CSWapp              -0.956  0.000  0.000 -0.169
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -3.05969653 -0.50019434 -0.02775304  0.44840513  2.96966986
##
## Number of Observations: 318
## Number of Groups: 106
```

```
Anova(ri.ed3)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: EATbelief
##            Chisq Df Pr(>Chisq)
## Timefact  1.6122  2  0.4465900
## Athlete  12.2262  1  0.0004712 ***
## CSWapp   14.9027  1  0.0001132 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Finally, while time does not have a significant directional effect and does not add significantly to our model based on Anova, we can refit our model to account for temporal dependence. Where each time point is dependent on the last.

```r
ri.ed4cor <- lme(EATbelief ~ Athlete + CSWapp,
             random = ~1|ID,
             data = ed.clean2,
             correlation = corAR1(form = ~Time))


summary(ri.ed4cor)
```

```
## Linear mixed-effects model fit by REML
##   Data: ed.clean2
##        AIC      BIC    logLik
##   580.6559 603.1713 -284.3279
##
## Random effects:
##  Formula: ~1 | ID
##         (Intercept)  Residual
## StdDev:   0.6060083 0.5333211
##
## Correlation Structure: AR(1)
##  Formula: ~Time | ID
##  Parameter estimate(s):
##       Phi
## 0.5018534
## Fixed effects:  EATbelief ~ Athlete + CSWapp
##                        Value Std.Error  DF  t-value p-value
## (Intercept)        0.7655079 0.3508011 212 2.182171  0.0302
## AthleteNon-Athlete 0.5068387 0.1448354 103 3.499413  0.0007
## CSWapp             0.2673441 0.0697563 103 3.832545  0.0002
##  Correlation:
##                    (Intr) AthN-A
## AthleteNon-Athlete -0.027
## CSWapp             -0.960 -0.169
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -2.17076098 -0.48923839 -0.09604347  0.31592622  2.87311182
##
## Number of Observations: 318
## Number of Groups: 106
```

This gives us the following model:

$EATbelief = .76 + .5NonAthlete + .26 * CSWappearence$

Interpretations and visualization

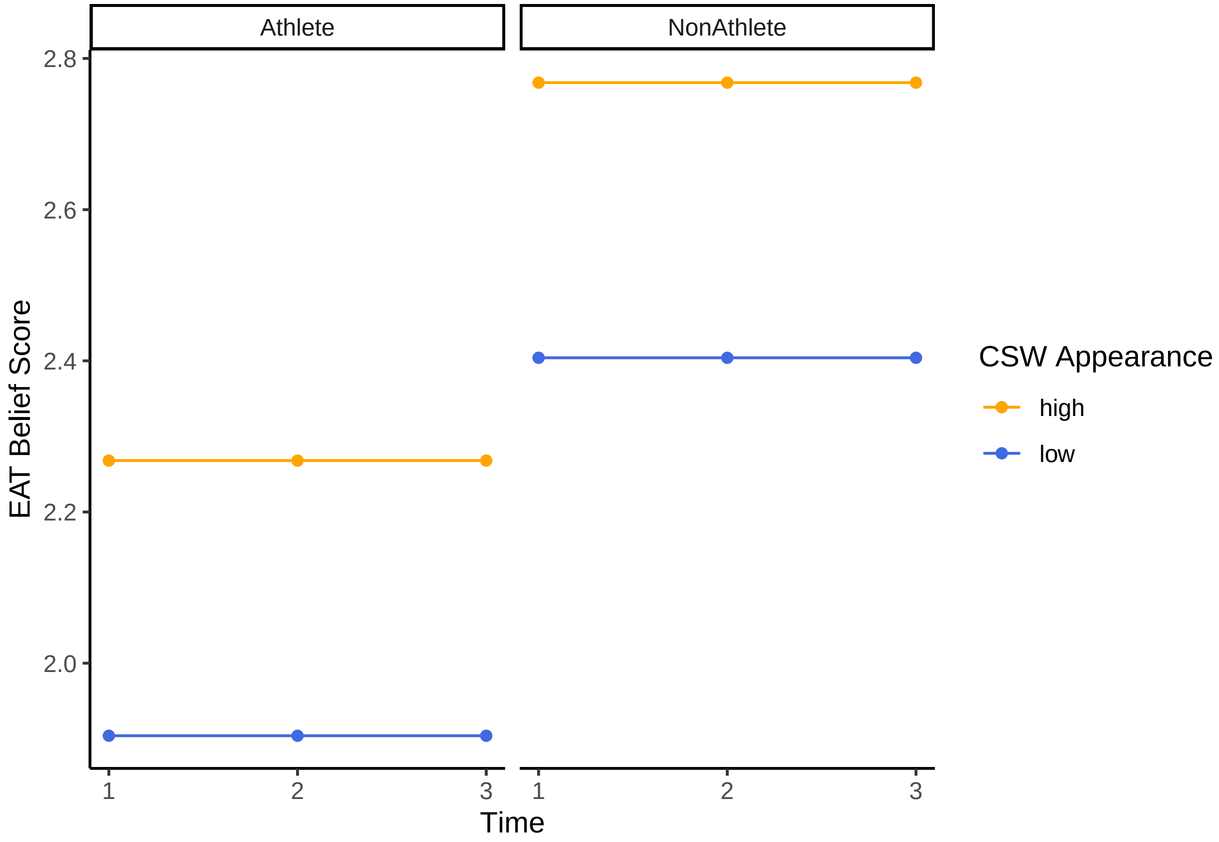
Based on the above model we have the following interpretations:

- We expect, on average, holding all else constant, that EATbelief scores will be .5 higher for Non-Athletes than for Athletes

- We expect, on average, holding all else constant, that EATbelief scores will increase by .26 for every 1-point increase in the Contingencies of Self Worth (Appearence).

Below is a visualization of our predicted values for athletes and non-athletes with high (3rd quartile) and low (1st quartile) CSWapp values

```
csw.sum <- summary(ed.clean2$CSWapp)

lowerq <- csw.sum[[2]]

upperq <- csw.sum[[5]]


new.data <- data.frame(Time = rep(c(1,2,3),4),
                       CSWapp = rep(c(lowerq, lowerq, lowerq,
                                      upperq, upperq, upperq), 2),
                       Athlete = as.factor(c(rep("NonAthlete", 6),
                                             rep("Athlete", 6))),
                       ID = rep(c(1, 2, 3, 4), each = 3),
                       Athlete01 = c(rep(1, 6), rep(0, 6)),
                       CSWapphl = rep(rep(c("low","high"), each = 3), 2))

new.data <- new.data %>%
  mutate(EATbelief = .76 + .5*Athlete01 + .26*CSWapp)

Final.Plot <- ggplot(new.data, aes(x = Time,
                                       y = EATbelief,
                                       color = CSWapphl,
                                       group = ID)) +
  geom_point() +
  geom_line() +
  facet_wrap(vars(Athlete)) +
  theme(legend.position = "top") +
  theme_classic()+
  labs(y = "EAT Belief Score") +
  scale_x_continuous(breaks = 1:3) +
  guides(color = guide_legend(title = "CSW Appearance")) +
  scale_color_manual(values = c("orange", "royalblue"))
Final.Plot
```



Conclusion (Brief)

While there is not a significant change over time, we can account for temporal dependence and see that both Non-Athlete status and CSWappearence scores have significant positive correlation with EATbelief scores.