combine\_158\_results

# Data Cleaning

## Load Data

get\_iat\_tibble <- function(file\_name, cp\_num){  
 df <- read.csv(file\_name) %>%  
 as\_tibble() %>%  
 select\_if(function(x) !(all(is.na(x)) | all(x==""))) %>%  
 na.omit() %>%  
 mutate(cp = cp\_num)  
 return(df)  
}  
  
aj1 <- get\_iat\_tibble('./RESULTS\_aj\_1.csv', 1)  
aj2 <- get\_iat\_tibble('./RESULTS\_aj\_2.csv', 1)

## Warning in read.table(file = file, header = header, sep = sep, quote  
## = quote, : incomplete final line found by readTableHeader on './  
## RESULTS\_aj\_2.csv'

rb1 <- get\_iat\_tibble('./RESULTS\_rb\_1.csv', 2)  
rb2 <- get\_iat\_tibble('./RESULTS\_rb\_2.csv', 2)

## Warning in read.table(file = file, header = header, sep = sep, quote  
## = quote, : incomplete final line found by readTableHeader on './  
## RESULTS\_rb\_2.csv'

rb3 <- get\_iat\_tibble('./RESULTS\_rb\_3.csv', 2)  
sd1 <- get\_iat\_tibble('./RESULTS\_sd\_1.csv', 3)  
sd2 <- get\_iat\_tibble('./RESULTS\_sd\_2.csv', 3)  
  
  
sd1 <- sd1 %>% rename\_at(vars(contains('X')), funs(sub('X', 'stage', .)))  
colnames(aj1) <- colnames(sd1)  
colnames(aj2) <- colnames(sd1)  
colnames(rb1) <- colnames(sd1)  
colnames(rb2) <- colnames(sd1)  
colnames(rb3) <- colnames(sd1)  
colnames(sd2) <- colnames(sd1)

## Remove extraneous observations so we have a balanced dataset

unbalanced <- rbind(aj1,aj2,rb1,rb2,rb3,sd1,sd2) %>%  
 mutate(time = chron(times = time))  
  
# We have an imbalanced dataset  
unbalanced %>%  
 group\_by(gender, factor) %>%  
 count()

## # A tibble: 8 x 3  
## # Groups: gender, factor [8]  
## gender factor n  
## <fct> <int> <int>  
## 1 Female 1 6  
## 2 Female 2 8  
## 3 Female 3 6  
## 4 Female 4 7  
## 5 Male 1 6  
## 6 Male 2 7  
## 7 Male 3 7  
## 8 Male 4 6

# Take only 6 subjects per treatment/block combination. If a treatment/block   
# combination k subjects, where k>6, remove the k-6 subjects who took  
# the experiment last. This is a neutral approach that does not meddle  
# with the randomization.  
balanced <- unbalanced %>%   
 group\_by(factor, gender) %>%  
 top\_n(n=6, wt=time) %>%  
 ungroup() %>%  
 arrange(time) %>%  
 mutate(subject\_id = 1:48)  
  
# Tidy the dataset, in the Hadley Wickham sense of 'tidy', not the Kon Mari sense  
tidy\_balanced <- balanced %>%  
 select(subject\_id, cp, time, factor, gender, everything()) %>%  
 gather(key = 'trial', value = 'response', 6:75)

## Remove outlier reaction times (not subjects)

tot\_responses <- nrow(tidy\_balanced)  
  
# Who has outlier response times? (i.e. response times that are too big)  
# in stage 3 and 5  
tidy\_balanced %>%  
 filter(!grepl("stage[1,2,4]",trial)) %>%  
 arrange(subject\_id) %>%  
 group\_by(subject\_id) %>%  
 filter(response > 10)

## # A tibble: 14 x 7  
## # Groups: subject\_id [10]  
## subject\_id cp time factor gender trial response  
## <int> <dbl> <S3: times> <int> <fct> <chr> <dbl>  
## 1 9 1 12:17:19 4 Male stage3\_0 11.3  
## 2 14 1 12:26:50 4 Male stage5\_0 10.1  
## 3 18 1 12:32:44 1 Male stage3\_15 21.4  
## 4 18 1 12:32:44 1 Male stage3\_16 10.9  
## 5 20 3 12:37:44 1 Male stage5\_0 20.2  
## 6 22 3 12:42:25 2 Female stage3\_1 10.3  
## 7 25 1 12:47:01 4 Male stage3\_1 11.3  
## 8 25 1 12:47:01 4 Male stage5\_3 11.3  
## 9 31 2 13:15:00 3 Female stage3\_0 11.4  
## 10 31 2 13:15:00 3 Female stage3\_9 42.8  
## 11 31 2 13:15:00 3 Female stage3\_19 30.8  
## 12 39 2 15:10:15 1 Female stage3\_1 10.8  
## 13 42 2 15:26:03 1 Female stage3\_0 10.0  
## 14 47 1 22:33:35 2 Male stage5\_4 12.1

# There are 14 outlier response times.  
  
# And there are 1920 total observations  
tidy\_balanced %>%  
 filter(!grepl("stage[1,2,4]",trial)) %>%  
 nrow()

## [1] 1920

# Remove outliers  
no\_outliers <- tidy\_balanced %>%  
 group\_by(subject\_id) %>%  
 filter(response < 10)  
  
# We don't need to remove anyone for cheating by clicking without reading  
# Greenwald only recommends removing data for participants where more than  
# 10% of subject response times are under 300 milliseconds. We have no such  
# subjects.  
no\_outliers %>%  
 group\_by(subject\_id) %>%  
 filter(response < .3) %>%  
 count()

## # A tibble: 6 x 2  
## # Groups: subject\_id [6]  
## subject\_id n  
## <int> <int>  
## 1 4 1  
## 2 5 1  
## 3 9 1  
## 4 11 1  
## 5 23 2  
## 6 38 2

## Feature Engineering: IAT Scores

# remove stages 1,2,4  
df <- no\_outliers %>%  
 filter(!grepl("stage[1,2,4]",trial)) %>%  
 arrange(subject\_id)  
  
  
# determine each subject's mean response time for  
# stage 3 (male\_leader) and 5 (female\_leader)  
mean\_response <- df %>%   
 mutate(male\_leader = grepl("stage3",trial)) %>%  
 group\_by(subject\_id, factor, gender, time, cp, male\_leader) %>%  
 summarise(mean = mean(response)) %>%  
 spread(key = male\_leader, value = mean) %>%  
 rename(stage3\_mean = 'TRUE', stage5\_mean = 'FALSE')  
  
  
# calculate iat score, which is the difference of mean response times for stage 3 & 5  
# divided by standard deviation of response times across both stages for that person  
iat\_df <- df %>%   
 group\_by(subject\_id) %>%  
 summarise(sd = sd(response)) %>%  
 right\_join(mean\_response) %>%  
 mutate(diff = (stage5\_mean - stage3\_mean)) %>%  
 mutate(iat\_d = diff/sd) %>%  
 select(-sd, sd) %>%  
 mutate(time = chron(times = time))

## Joining, by = "subject\_id"

# Add a binary feature for the factor. Helps for ANOVA.  
iat\_df <- iat\_df %>%  
 mutate(prodiversity = (factor %in% c(2,4))) %>%  
 mutate(antistereotype = (factor %in% c(3,4)))

# Write out csv & data table

write.csv(iat\_df, file = './iat\_scores.csv')  
  
pretty\_table <- iat\_df %>%  
 select(subject\_id, time, gender, cp, prodiversity, antistereotype, iat\_d) %>%  
 mutate(iat\_d = round(iat\_d,3))  
  
write.csv(pretty\_table, file = './iat\_scores\_table.csv')

# Exploratory Analysis

## Checking Greenwald’s recommendation that we

# People with more variable reaction times tended to have larger differences between stage 3 and 5  
cor(abs(iat\_df$diff), iat\_df$sd)

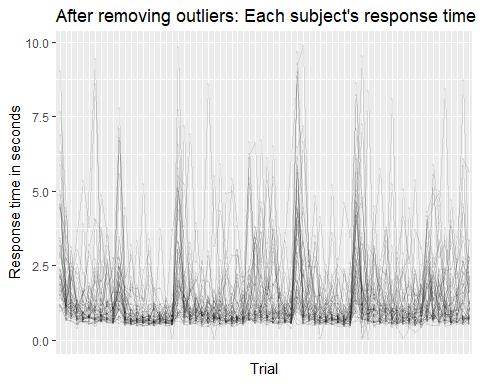
## [1] 0.5713127

This confirms Greenwald’s claim that there is great association between sd of reaction times and the absolute value of the difference in mean reaction times for stage 3 and 5.

## Plot reaction times from stages 1 through 5

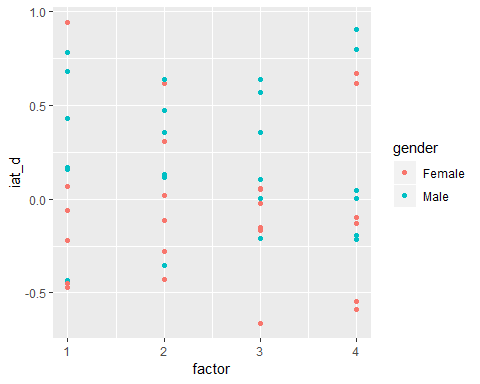
# As we expect, reaction times jump up when a new stage begins.  
ggplot(no\_outliers, aes(trial,response, group=time)) +  
 geom\_point(alpha=.05, size=.01)+  
 geom\_line(alpha=.1, size=.1)+  
 xlab("Trial")+   
 ylab("Response time in seconds") +  
 ggtitle("After removing outliers: Each subject's response time during the experiment")+  
 theme(axis.text.x=element\_blank(),  
 axis.ticks.x=element\_blank())

## Don't know how to automatically pick scale for object of type times. Defaulting to continuous.  
## Don't know how to automatically pick scale for object of type times. Defaulting to continuous.



# Blocking

# Blocking by gender was a good call  
ggplot(iat\_df) +  
 geom\_point(aes(x=factor, y = iat\_d, color = gender))



iat\_df

## # A tibble: 48 x 12  
## subject\_id factor gender time cp stage5\_mean stage3\_mean diff  
## <int> <int> <fct> <S3:> <dbl> <dbl> <dbl> <dbl>  
## 1 1 3 Male 10:3~ 2 1.35 1.23 0.117   
## 2 2 3 Female 10:3~ 2 0.903 1.02 -0.122   
## 3 3 2 Female 10:4~ 2 2.11 1.65 0.459   
## 4 4 4 Female 10:4~ 2 2.33 1.25 1.08   
## 5 5 1 Male 10:5~ 1 1.22 1.06 0.155   
## 6 6 4 Male 11:4~ 3 1.03 1.00 0.0292  
## 7 7 4 Female 12:0~ 3 0.760 1.04 -0.280   
## 8 8 3 Male 12:1~ 3 1.27 0.876 0.394   
## 9 9 4 Male 12:1~ 1 2.12 1.05 1.07   
## 10 10 1 Female 12:1~ 3 1.50 1.42 0.0742  
## # ... with 38 more rows, and 4 more variables: iat\_d <dbl>, sd <dbl>,  
## # prodiversity <lgl>, antistereotype <lgl>

## Contrast: Treatments vs Control

Let’s examine the contrast for whether there was a difference in response between the control group and the treatment groups. This corresponds to the coefficient vector of (-3, -3, 1, 1, 1, 1, 1, 1). In that vector, the first two entries correspond to the two control groups (one per gender); the remaining four entries correspond to the remaining treatment / block combinations.

iat\_df %>%  
 group\_by(factor, gender) %>%  
 summarise(mean(diff))

## # A tibble: 8 x 3  
## # Groups: factor [?]  
## factor gender `mean(diff)`  
## <int> <fct> <dbl>  
## 1 1 Female -0.0421  
## 2 1 Male 0.273   
## 3 2 Female 0.122   
## 4 2 Male 0.188   
## 5 3 Female -0.216   
## 6 3 Male 0.247   
## 7 4 Female 0.127   
## 8 4 Male 0.280

Thus, the unbiased estimate is , which equals 0.0571. We use MS residuals as an estimate for error, by multiplying it by:

3\*(.0421) -3\*(.2727) + .1218 + .1884 - .2159 + .2474 + .1271 + .2801

## [1] 0.0571

(-3)^2/6 + (-3)^2/6 + 1/6 + 1/6 + 1/6 + 1/6 + 1/6

## [1] 3.833333

Our t-statistic then is: 0.0571/sqrt(.1727 \* 3.833333)

0.0571/sqrt(.1691 \* 3.833333)

## [1] 0.07092119

(1-pt(.07, df = 48-6))\*2

## [1] 0.9445259

summary(aov(iat\_d ~ prodiversity + antistereotype + prodiversity:antistereotype + gender, data = iat\_df))

## Df Sum Sq Mean Sq F value Pr(>F)   
## prodiversity 1 0.007 0.0072 0.041 0.8396   
## antistereotype 1 0.031 0.0315 0.182 0.6717   
## gender 1 1.019 1.0191 5.900 0.0194 \*  
## prodiversity:antistereotype 1 0.014 0.0141 0.082 0.7762   
## Residuals 43 7.428 0.1727   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

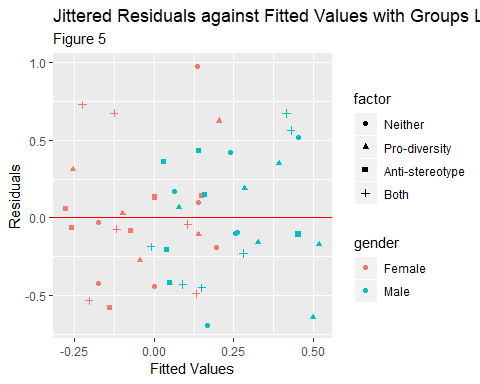
## ANOVA Assumptions:

## Residual Check

# Fitted vs. Residual  
mod = aov(iat\_d ~ prodiversity + antistereotype + gender, data = iat\_df)  
  
fit\_resid = tibble(f = fitted(mod), r = residuals(mod))  
  
fit\_resid

## # A tibble: 48 x 2  
## f r  
## <dbl> <dbl>  
## 1 0.211 -0.107   
## 2 -0.0808 -0.0857  
## 3 -0.00516 0.313   
## 4 -0.0564 0.728   
## 5 0.262 -0.102   
## 6 0.235 -0.186   
## 7 -0.0564 -0.489   
## 8 0.211 0.430   
## 9 0.235 0.563   
## 10 -0.0296 0.0952  
## # ... with 38 more rows

iat\_df <- mutate(iat\_df, factor = as.factor(factor))  
  
levels(iat\_df$factor) <- c('Neither', 'Pro-diversity','Anti-stereotype','Both')  
ggplot(iat\_df, aes(x=fit\_resid$f, y=fit\_resid$r)) +   
 geom\_jitter(width = 0.25,aes(color = gender, shape = factor)) +   
 geom\_hline(yintercept = 0, color='red') +   
 ggtitle("Jittered Residuals against Fitted Values with Groups Labelled") +   
 xlab("Fitted Values") +   
 ylab("Residuals")+  
 labs(subtitle = "Figure 5")



model.tables(mod)

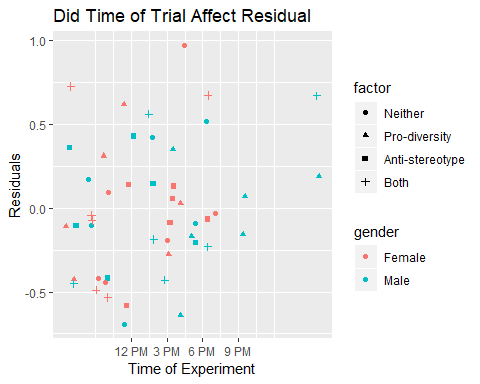
## Warning in replications(paste("~", xx), data = mf): non-factors ignored:  
## prodiversity

## Warning in replications(paste("~", xx), data = mf): non-factors ignored:  
## antistereotype

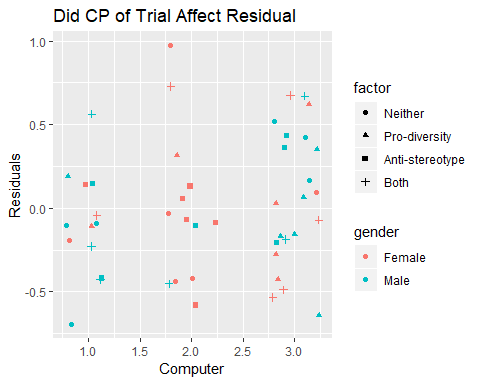
## Tables of effects  
##   
## prodiversity   
## prodiversity  
## FALSE TRUE   
## -0.012217 0.012217   
##   
## antistereotype   
## antistereotype  
## FALSE TRUE   
## 0.025598 -0.025598   
##   
## gender   
## gender  
## Female Male   
## -0.14571 0.14571

### Independence

# Is response time determined by when they took the IAT  
ggplot(iat\_df, aes(x=time, y=fit\_resid$r)) +   
 geom\_jitter(width = 0.25,aes(color = gender, shape = factor)) +   
 ylab("Residuals") +  
 xlab("Time of Experiment") +  
 ggtitle('Did Time of Trial Affect Residual')+  
 scale\_x\_continuous(breaks = c(.5,.625, .75, .875),   
 labels = c('12 PM', '3 PM', '6 PM','9 PM'))



# Is response time determined by the computer on which they took the IAT  
ggplot(iat\_df, aes(x=cp, y=fit\_resid$r)) +   
 geom\_jitter(width = 0.25,aes(color = gender, shape = factor)) +   
 ylab("Residuals") +  
 xlab("Computer") +  
 ggtitle('Did CP of Trial Affect Residual')



Neither assumption is violated.