MANOVA-equalcovars

2022-08-25

FROM ANN:

For MANOVA (= formal analysis), may want to think about how to use “other” or “no answer” categories. Also consider whether you want to include all interactions. If you do, consider the combination of the two issues above.

When retaining “other” and “no answer” categories, there are a total of 3 genders (including 14 Other) 5 races (including 29 Other) 3 age groups This yields a total of 3*5*3 = 45 possible combinations, but only observe 35 total combinations. Some of those combinations include only 1-2 people.

ADDITIVE MODEL WAS SUGGESTED BY ANN

Use summary() to get tests corresponding to predictors. For MANOVA there are several possible methods for testing.

Personally, I would not worry about doing Box’s test. You might mention the corresponding assumption, but I am not inclined to do a formal test. Or if you really do want to run it, perhaps do it separately by each factor (gender, race, age\_group).

MY NOTES:

To check normality: QQ plot (A Q-Q plot or Quantile-Quantile plot is a common graphical way to check data for non-normality. Quantile is another term for percentile. A Q-Q plot is a plot of the quantiles of a data set versus the quantiles of a reference theoretical distribution. This plot is used to assess normality of residuals. Curvature and outliers indicate problems. Note: The QQ plot is not useful until variance is approximately equal.)

shapiro.test with p value greater than 0.05 (Histograms and QQ plots are usually more informative than the tests, because small sample sizes generally “pass” the test (high p-value, no evidence against normality), and large sample sizes generally “fail” (small p-value, evidence against normality).)

To check homogeneity of covariances/equal variance:

Box’s test (homegeneity of variance-covariance matrices)

Levene’s test of equality of variances (homogeneity of variances)

Plot of residuals vs fitted values (equal scatter; This is the primary diagnostic plot for assessing linearity and constant variance. Curvature, unequal variance (megaphone) or outliers indicate problems.)

To check linear response:

Plot of resids vs fitted values (should not show a trend)

Look to pages 20-end 512 MultReg1

For y variables: You should have an adequate sample size. Although the larger your sample size, the better, at a bare minimum, there needs to be as many cases (e.g., particpants) in each cell of the design as there are number of dependent variables. 12 dependent variables

For MANOVA, the dependent variables should be somewhat correlated.

STEPS: Power analysis to find sample size - 278 NAs shouldn’t be over 20% Find outliers & remove (see removing stat outliers or MANOVA imputing) Double check correlations check assumptions run manova run anovas

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4  
## ✔ tibble 3.1.8 ✔ dplyr 1.0.9  
## ✔ tidyr 1.2.0 ✔ stringr 1.4.0  
## ✔ readr 2.1.2 ✔ forcats 0.5.1  
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(readxl)  
library(flextable)

##   
## Attaching package: 'flextable'  
##   
## The following object is masked from 'package:purrr':  
##   
## compose

library(extrafont)

## Registering fonts with R

library(writexl)  
library(corrr)  
library(corrplot)

## corrplot 0.92 loaded

library(heplots) #boxM function

## Loading required package: car  
## Loading required package: carData  
##   
## Attaching package: 'car'  
##   
## The following object is masked from 'package:dplyr':  
##   
## recode  
##   
## The following object is masked from 'package:purrr':  
##   
## some

library(DFA.CANCOR)

## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
## DFA.CANCOR 0.2.5  
##   
## Please contact Brian O'Connor at brian.oconnor@ubc.ca if you have questions or suggestions.  
## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

#times new roman tables  
my\_ft\_theme <- function(ft, ...) {  
 # Remove vertical cell padding  
 ft <- padding(ft, padding.top = 0, padding.bottom = 0, part = "all")  
   
 # Change font to TNR 11  
 ft <- font(ft, fontname = "Times New Roman", part = "all")  
 ft <- fontsize(ft, part = "all", size = 12)  
 ft  
}

SA3\_noout <- read\_excel("C:\\Users\\19177\\OneDrive - Colostate\\Desktop\\Dissertation\\headscan\_dissertation\\chosen\_nona1.xlsx")

str(SA3\_noout)

## tibble [1,677 × 16] (S3: tbl\_df/tbl/data.frame)  
## $ ID : chr [1:1677] "400-20201012-002" "400-20201012-003" "400-20201012-004" "400-20201012-005" ...  
## $ AA\_C : num [1:1677] 65 55 70 58 67 60 59 59 65 65 ...  
## $ BiW\_C : num [1:1677] 130 127 143 140 137 130 141 138 143 150 ...  
## $ BiW\_L : num [1:1677] 115 108 121 109 104 106 109 111 113 116 ...  
## $ GoSub\_C : num [1:1677] 93 93 115 93 103 100 79 106 85 102 ...  
## $ NRB\_L : num [1:1677] 17 18 19 21 19 14 17 18 16 17 ...  
## $ ProS\_L : num [1:1677] 17 18 14 13 20 20 18 12 24 22 ...  
## $ SelP\_L : num [1:1677] 42 41 51 44 47 48 46 41 46 44 ...  
## $ SelM\_L : num [1:1677] 122 99 130 115 119 126 117 112 117 117 ...  
## $ SnasM\_C : num [1:1677] 82 55 84 74 73 80 78 76 64 75 ...  
## $ TrSman\_C : num [1:1677] 177 145 178 147 157 164 149 159 151 160 ...  
## $ TrTr\_C : num [1:1677] 296 276 292 273 279 300 283 275 307 286 ...  
## $ TrTr\_L : num [1:1677] 155 141 156 149 146 146 147 151 157 144 ...  
## $ gender : chr [1:1677] "Male" "Female" "Male" "Male" ...  
## $ race\_eth : chr [1:1677] "Black" "white" "white" "white" ...  
## $ age\_group: chr [1:1677] "18-36" "37-54" "37-54" "18-36" ...

SA3\_num <- select\_if(SA3\_noout, is.numeric)

correl <- correlate(SA3\_num)

##   
## Correlation method: 'pearson'  
## Missing treated using: 'pairwise.complete.obs'

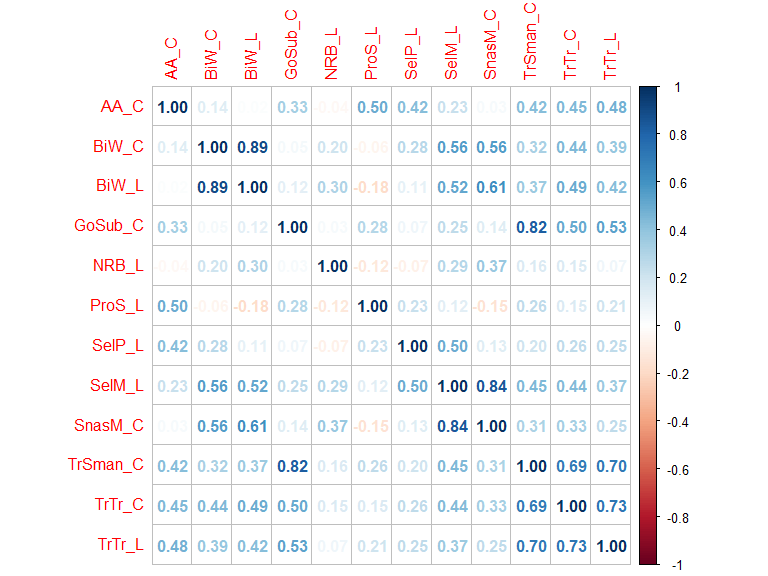
correl

## # A tibble: 12 × 13  
## term AA\_C BiW\_C BiW\_L GoSub\_C NRB\_L ProS\_L SelP\_L SelM\_L SnasM\_C  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 AA\_C NA 0.140 0.0153 0.334 -0.0395 0.497 0.423 0.227 0.0311  
## 2 BiW\_C 0.140 NA 0.892 0.0494 0.205 -0.0594 0.280 0.555 0.557   
## 3 BiW\_L 0.0153 0.892 NA 0.117 0.302 -0.182 0.106 0.516 0.613   
## 4 GoSub… 0.334 0.0494 0.117 NA 0.0330 0.282 0.0682 0.254 0.137   
## 5 NRB\_L -0.0395 0.205 0.302 0.0330 NA -0.120 -0.0740 0.286 0.371   
## 6 ProS\_L 0.497 -0.0594 -0.182 0.282 -0.120 NA 0.227 0.120 -0.154   
## 7 SelP\_L 0.423 0.280 0.106 0.0682 -0.0740 0.227 NA 0.497 0.135   
## 8 SelM\_L 0.227 0.555 0.516 0.254 0.286 0.120 0.497 NA 0.844   
## 9 SnasM… 0.0311 0.557 0.613 0.137 0.371 -0.154 0.135 0.844 NA   
## 10 TrSma… 0.423 0.320 0.367 0.822 0.159 0.255 0.202 0.445 0.312   
## 11 TrTr\_C 0.446 0.442 0.490 0.503 0.151 0.154 0.257 0.439 0.334   
## 12 TrTr\_L 0.477 0.389 0.421 0.535 0.0667 0.205 0.253 0.368 0.252   
## # … with 3 more variables: TrSman\_C <dbl>, TrTr\_C <dbl>, TrTr\_L <dbl>  
## # ℹ Use `colnames()` to see all variable names

correl2 <- cor(SA3\_num)  
symnum(correl2)

## A BW\_C BW\_L G N P SP SM\_L SM\_C TS TT\_C TT\_L  
## AA\_C 1   
## BiW\_C 1   
## BiW\_L + 1   
## GoSub\_C . 1   
## NRB\_L . 1   
## ProS\_L . 1   
## SelP\_L . 1   
## SelM\_L . . . 1   
## SnasM\_C . , . + 1   
## TrSman\_C . . . + . . 1   
## TrTr\_C . . . . . . , 1   
## TrTr\_L . . . . . , , 1   
## attr(,"legend")  
## [1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '\*' 0.95 'B' 1

corrplot(correl2, method = 'number')



Checking assumptions on actual MANOVA model: setup

variables <- cbind(SA3\_noout$AA\_C, SA3\_noout$BiW\_C, SA3\_noout$BiW\_L,   
 SA3\_noout$GoSub\_C, SA3\_noout$NRB\_L, SA3\_noout$ProS\_L,  
 SA3\_noout$SelP\_L, SA3\_noout$SelM\_L, SA3\_noout$SnasM\_C,  
 SA3\_noout$TrSman\_C, SA3\_noout$TrTr\_C, SA3\_noout$TrTr\_L)  
  
  
#part of Dr. Erin Buchanan's lecture, but Manova code returns error: Error in Anova.mlm(mod, ...) : model is singular  
#lm\_notransf <- lm(variables ~ gender+race\_eth+age\_group, data=SA3\_noout,  
 #contrasts = list(gender=contr.sum,   
 #race\_eth=contr.sum,  
 #age\_group=contr.sum))  
  
#manova\_notransf <- Manova(lm\_notransf, type = "III")

Checking assumptions on actual MANOVA model: no transformation

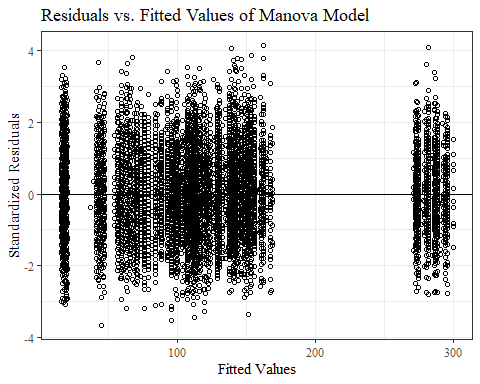
manova\_out\_nt <- manova(variables ~ gender+race\_eth+age\_group, data=SA3\_noout)  
fitted1 <- manova\_out\_nt$fitted.values  
resid1 <- rstandard(manova\_out\_nt)  
  
#homogeneity (unscaled)  
#plot(fitted1, resid1)  
#abline(0,0)  
#abline(v = 0)  
  
#linearity  
#qqnorm(resid1)  
#abline(0,1, lty=2)  
  
#normality  
#hist(resid1)

#creating dataframe for ggplot  
fitted1.1 <- as.data.frame(fitted1)  
resid1.1 <-as.data.frame(resid1)  
  
fitted1.1 <- rownames\_to\_column(fitted1.1, "scan\_num")  
  
fitted1.1 <- pivot\_longer(fitted1.1, V1:V12, names\_to = "variable", values\_to = "fitted\_vals")  
  
  
resid1.1 <- rownames\_to\_column(resid1.1, "scan\_num")  
  
resid1.1 <- pivot\_longer(resid1.1, V1:V12, names\_to = "variable", values\_to = "resid\_vals")  
  
  
  
temp5 <- full\_join(fitted1.1, resid1.1, by=c('scan\_num', 'variable'))  
str(temp5)

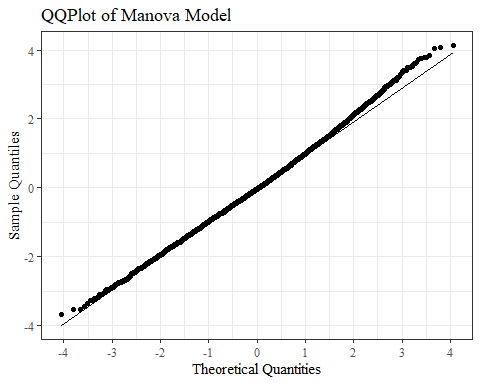
## tibble [20,124 × 4] (S3: tbl\_df/tbl/data.frame)  
## $ scan\_num : chr [1:20124] "1" "1" "1" "1" ...  
## $ variable : chr [1:20124] "V1" "V2" "V3" "V4" ...  
## $ fitted\_vals: num [1:20124] 64.1 139 116.5 104 19.4 ...  
## $ resid\_vals : num [1:20124] 0.165 -0.76 -0.161 -0.863 -0.528 ...

THIS is the model for reporting, and for use in MANOVA.

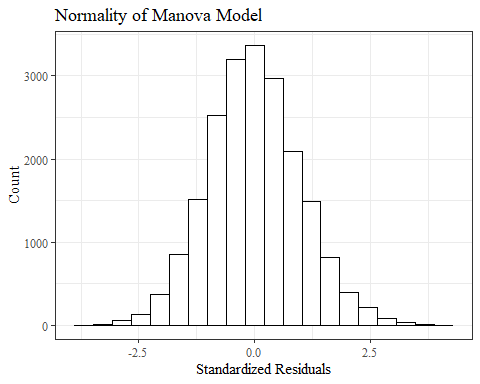
ggplot(data=temp5, aes(x=fitted\_vals, y=resid\_vals))+  
 geom\_point(shape=1)+  
 geom\_hline(yintercept=0)+   
 theme\_bw()+  
 theme(text=element\_text(family= "Times New Roman"))+  
 labs(title="Residuals vs. Fitted Values of Manova Model",  
 y= "Standardized Residuals",  
 x= "Fitted Values")



ggplot(data=temp5, aes(sample=resid\_vals))+  
 stat\_qq()+  
 stat\_qq\_line()+  
 scale\_x\_continuous(n.breaks = 7)+  
 #xlim(-3.5,3.5)+  
 theme\_bw()+  
 theme(text=element\_text(family= "Times New Roman"))+  
 labs(title="QQPlot of Manova Model",  
 y= "Sample Quantiles",  
 x= "Theoretical Quantities")



ggplot(data=temp5, aes(x=resid\_vals))+  
 geom\_bar(stat="bin", bins=20, color= "black", fill = "white")+  
 theme\_bw()+theme(text=element\_text(family= "Times New Roman"))+  
 labs(title="Normality of Manova Model",  
 y= "Count",  
 x= "Standardized Residuals")



boxM( cbind(AA\_C, BiW\_C, BiW\_L, GoSub\_C, NRB\_L, ProS\_L,  
 SelP\_L, SelM\_L, SnasM\_C,   
 TrSman\_C, TrTr\_C, TrTr\_L) ~ race\_eth, data=SA3\_noout)

##   
## Box's M-test for Homogeneity of Covariance Matrices  
##   
## data: Y  
## Chi-Sq (approx.) = 548.89, df = 312, p-value = 2.797e-15

boxM( cbind(AA\_C, BiW\_C, BiW\_L, GoSub\_C, NRB\_L, ProS\_L,  
 SelP\_L, SelM\_L, SnasM\_C,  
 TrSman\_C, TrTr\_C, TrTr\_L) ~ gender, data=SA3\_noout)

## Warning in boxM.default(Y = Y, group = group, ...): there are one or more levels  
## with less observations than variables!

##   
## Box's M-test for Homogeneity of Covariance Matrices  
##   
## data: Y  
## Chi-Sq (approx.) = -Inf, df = 156, p-value = 1

boxM( cbind(AA\_C, BiW\_C, BiW\_L, GoSub\_C, NRB\_L, ProS\_L,  
 SelP\_L, SelM\_L, SnasM\_C,  
 TrSman\_C, TrTr\_C, TrTr\_L) ~ age\_group, data=SA3\_noout)

##   
## Box's M-test for Homogeneity of Covariance Matrices  
##   
## data: Y  
## Chi-Sq (approx.) = 206.76, df = 156, p-value = 0.004055

SA3\_noout$gender <- as.factor(SA3\_noout$gender)  
SA3\_noout$race\_eth <- as.factor(SA3\_noout$race\_eth)  
SA3\_noout$age\_group <- as.factor(SA3\_noout$age\_group)  
  
summary(SA3\_noout)

## ID AA\_C BiW\_C BiW\_L   
## Length:1677 Min. :44.00 Min. :101.0 Min. : 82.0   
## Class :character 1st Qu.:56.00 1st Qu.:123.0 1st Qu.:104.0   
## Mode :character Median :60.00 Median :131.0 Median :110.0   
## Mean :60.89 Mean :132.4 Mean :110.5   
## 3rd Qu.:65.00 3rd Qu.:141.0 3rd Qu.:117.0   
## Max. :87.00 Max. :187.0 Max. :148.0   
## GoSub\_C NRB\_L ProS\_L SelP\_L   
## Min. : 49.00 Min. : 3.00 Min. :12.00 Min. :31.00   
## 1st Qu.: 88.00 1st Qu.:15.00 1st Qu.:17.00 1st Qu.:42.00   
## Median : 97.00 Median :18.00 Median :19.00 Median :44.00   
## Mean : 97.87 Mean :17.86 Mean :19.09 Mean :44.33   
## 3rd Qu.:107.00 3rd Qu.:21.00 3rd Qu.:21.00 3rd Qu.:47.00   
## Max. :152.00 Max. :34.00 Max. :27.00 Max. :58.00   
## SelM\_L SnasM\_C TrSman\_C TrTr\_C   
## Min. : 84.0 Min. : 46.00 Min. :117.0 Min. :241.0   
## 1st Qu.:110.0 1st Qu.: 68.00 1st Qu.:142.0 1st Qu.:271.0   
## Median :116.0 Median : 75.00 Median :151.0 Median :281.0   
## Mean :116.3 Mean : 74.96 Mean :152.2 Mean :281.2   
## 3rd Qu.:123.0 3rd Qu.: 82.00 3rd Qu.:161.0 3rd Qu.:291.0   
## Max. :145.0 Max. :105.00 Max. :208.0 Max. :329.0   
## TrTr\_L gender race\_eth age\_group   
## Min. :127.0 Female:994 Asian : 81 18-36:826   
## 1st Qu.:140.0 Male :681 Black : 446 37-54:777   
## Median :145.0 Other : 2 LatinX: 84 55-72: 74   
## Mean :145.7 Other : 26   
## 3rd Qu.:151.0 white :1040   
## Max. :172.0

boxtest\_data <- SA3\_noout  
  
boxtest\_data$gender <-  
 recode\_factor(boxtest\_data$gender, 'Female'= "Female/Other",  
 'Other' = "Female/Other")  
  
boxM( cbind(AA\_C, BiW\_C, BiW\_L, GoSub\_C, NRB\_L, ProS\_L,  
 SelP\_L, SelM\_L, SnasM\_C,  
 TrSman\_C, TrTr\_C, TrTr\_L) ~ gender, data=boxtest\_data)

##   
## Box's M-test for Homogeneity of Covariance Matrices  
##   
## data: Y  
## Chi-Sq (approx.) = 262.06, df = 78, p-value < 2.2e-16

Box’s M test the assumption of equality of covariance matrices. However, Box’s M is sensitive to large data files; therefore, it can detect even small departures from homogeneity. Moreover, it can be sensitive to departures from the assumption of normality.

Box’s M: If group sizes are over 30, then the MANOVA is robust against violations of homogeneity of variance-covariance matrices assumption. Levene’s Test: if homogeneity of variance can’t be assumed for one (or more) dependent variables, then use an alpha level stricter than .05 (ie: use .001) when you evaluate the univariate ANOVAs. [From “SPSS for the Health and Behavioural Sciences” (Allen & Bennett, 2008)]

Huberty, C. J, & Morris, J. D. (1989). Multivariate analysis versus multiple univariate analyses. Psychological Bulletin, 105, 302-308.

<https://www.researchgate.net/post/What-if-the-Boxs-Test-in-two-way-MANOVA-is-significant>

#don’t really see the point of doing transformations, they didn’t affect too much for visual assumptions or Levene’s test

#{r} HOMOGENEITY(data=SA3\_noout, groups='race\_eth', variables=c('AA\_C', 'BiW\_C', 'BiW\_L', 'GoSub\_C', 'NRB\_L', 'ProS\_L', 'SelP\_L', 'SelM\_L', 'SnasM\_C','TrSman\_C', 'TrTr\_C', 'TrTr\_L')) # #{r} HOMOGENEITY(data=SA3\_noout, groups='gender', variables=c('AA\_C', 'BiW\_C', 'BiW\_L', 'GoSub\_C', 'NRB\_L', 'ProS\_L', 'SelP\_L', 'SelM\_L', 'SnasM\_C', 'TrSman\_C', 'TrTr\_C', 'TrTr\_L')) # #{r} HOMOGENEITY(data=SA3\_noout, groups='age\_group', variables=c('AA\_C', 'BiW\_C', 'BiW\_L', 'GoSub\_C', 'NRB\_L', 'ProS\_L', 'SelP\_L', 'SelM\_L', 'SnasM\_C', 'TrSman\_C', 'TrTr\_C', 'TrTr\_L')) #

#```{r} boxtest\_data <- column\_to\_rownames(SA3\_noout, “ID”)

boxtest\_datarace\_eth, ‘LatinX’= “Other”, ‘Asian’ = “Other”)

boxtest\_datagender, ‘Female’= “Female or Other”, ‘Other’ = “Female or Other”)

boxtest\_data <- boxtest\_data %>% drop\_na()

boxtest\_dataage\_group)

summary(boxtest\_data)

boxM( cbind(AA\_C, BiW\_C, BiW\_L, GoSub\_C, NRB\_L, ProS\_L, SelP\_L, SelM\_L, SnasM\_C, TrSman\_C, TrTr\_C, TrTr\_L) ~ race\_eth, data=boxtest\_data)

boxM( cbind(AA\_C, BiW\_C, BiW\_L, GoSub\_C, NRB\_L, ProS\_L, SelP\_L, SelM\_L, SnasM\_C, TrSman\_C, TrTr\_C, TrTr\_L) ~ gender, data=boxtest\_data)

boxM( cbind(AA\_C, BiW\_C, BiW\_L, GoSub\_C, NRB\_L, ProS\_L, SelP\_L, SelM\_L, SnasM\_C, TrSman\_C, TrTr\_C, TrTr\_L) ~ age\_group, data=boxtest\_data)

##{r} HOMOGENEITY(data=boxtest\_data, groups=‘race\_eth’, variables=c(‘AA\_C’, ‘BiW\_C’, ‘BiW\_L’, ‘GoSub\_C’, ‘NRB\_L’, ‘ProS\_L’, ‘SelP\_L’, ‘SelM\_L’, ‘SnasM\_C’, ‘TrSman\_C’, ‘TrTr\_C’, ‘TrTr\_L’)) #```

#{r} HOMOGENEITY(data=boxtest\_data, groups='gender', variables=c('AA\_C', 'BiW\_C', 'BiW\_L', 'GoSub\_C', 'NRB\_L', 'ProS\_L', 'SelP\_L', 'SelM\_L', 'SnasM\_C', 'TrSman\_C', 'TrTr\_C', 'TrTr\_L')) #

#{r} HOMOGENEITY(data=boxtest\_data, groups='age\_group', variables=c('AA\_C', 'BiW\_C', 'BiW\_L', 'GoSub\_C', 'NRB\_L', 'ProS\_L', 'SelP\_L', 'SelM\_L', 'SnasM\_C', 'TrSman\_C', 'TrTr\_C', 'TrTr\_L')) #

Box’s M test the assumption of equality of covariance matrices. However, Box’s M is sensitive to large data files; therefore, it can detect even small departures from homogeneity. Moreover, it can be sensitive to departures from the assumption of normality.

Box’s M: If group sizes are over 30, then the MANOVA is robust against violations of homogeneity of variance-covariance matrices assumption. Levene’s Test: if homogeneity of variance can’t be assumed for one (or more) dependent variables, then use an alpha level stricter than .05 (ie: use .001) when you evaluate the univariate ANOVAs. [From “SPSS for the Health and Behavioural Sciences” (Allen & Bennett, 2008)]

Huberty, C. J, & Morris, J. D. (1989). Multivariate analysis versus multiple univariate analyses. Psychological Bulletin, 105, 302-308.