BIOSTATS 640 – Introduction to R
Fall 2023
https://people.umass.edu/biep640w/webpages/demonstrations.html

https://sixsigmastudyguide.com/anova-analysis-of-variation/

09
Introduction to Analysis of Variance
November 3, 2023

Dataset used
hers_640anova.xlsx

		Page
1	Introduction to The Heart and Estrogen/Progestin Replacement Study (HERS): **hers_640anova.xlsx**	2
2	Highlights of Lesson 08 – Regression Diagnostics for Normal Theory Regression	3
3.	**One Way Analysis of Variance** ..... 3.1. Prepare Data: Create Factors and Set Reference ..... 3.2 Data Description ..... 3.3 Model Estimation and Interpretation ..... 3.4 Regression Diagnostics ..... 3.5 Post-hoc Tests and Estimation ..... 3.6 Post-hoc Visualizations	4 4 4 8 10 11 13
4.	**Two Way Factorial Analysis of Variance** ..... 4.1. Prepare Data: Create Factors and Set References ..... 4.2 Data Description ..... 4.3 Model Estimation and Interpretation ..... 4.4 Post-hoc Tests and Estimation ..... 4.5 Post-hoc Visualizations	15 15 16 19 21 25

Packages used: `ggplot2, summarytools, knitr, tidyverse, HH, car, multcomp, FSA, emmeans, gridExtra, MASS`

1. Introduction to The Heart and Estrogen/progestin Replacement Study (HERS)

hers_640anova.xlsx

Source

Hulley et al (1998) Randomized trial of estrogen plus progestin for secondary prevention of heart disease in postmenopausal women. The Heart and Estrogen/progestin Replacement Study. *Journal of the American Medical Association*, **280**(7), 605-613

The Heart and Estrogen/Progestin Replacement Study (HERS) was a randomized clinical trial of hormone therapy (estrogen plus progestin) for the reduction of cardiovascular disease risk in post-menopausal women with established coronary disease. Study participants were n=2,763 women who were: (1) post-menopausal (2) with coronary disease; and (3) with an intact uterus.

This illustration uses a subset of the data with n = 612. Three variables are considered:

Data dictionary/Codebook *(Partial)*

Variable	Label	Type	Codings
sbp	Systolic blood pressure (mm Hg)	numeric	Continuous, range, [45:79]
raceth	Race	numeric	1 = White 2 = African American 3 = Other
physact	Comparative ("compared to other women your age") physical activity	Numeric	1 = much less active 2 = somewhat less active 3 = about as active 4 = somewhat more active 5 = much more active

2. Highlights of Lesson 08
Regression Diagnostics for Normal Theory Regression in R

plot( ). No package necessary

Command	Plot Produced
plot(fit, which=1)	X = fitted value Y = residual
plot(fit, which=2)	X = theoretical normal quantile Y = studentized residual
plot(fit, which=3)	X = fitted value Y = square root (standardized residual)
plot(fit, which=4)	X = observation number Y = Cook's Distance
plot(fit, which=5)	X = leverage Y = standardized residual
plot(fit, which=6)	X = leverage Y = Cook's Distance
plot(fit)	Default is four plots: which=1, which=2, which=3, and which=5

residualPlots( ) in package {car}

This command also provides,for each predictor X, a t-test of NULL: "no curvature" quadratic X^2 is not statistically significant. It also provides the Tukey test of NULL: "the model is additive"

Command	Plots Produced
residualPlots(fit)	For each predictor: X = predictor Y = residual And also: X = fitted Y = residual
residualPlots(fit, ~X1)	For single predictor of interest: X = predictor Y = residual And also: X = fitted Y = residual
residualPlots(fit, ~1)	X = fitted Y = residual *ONLY*

autoplot( ) in package {ggfortify}.

To be safe you might need to have library(ggplot2)

Command	Plot Produced
autoplot(fit, which=1, option, option)	X = fitted value Y = residual
autoplot(fit, which=2, option, option)	X = theoretical normal quantile Y = studentized residual
autoplot(fit, which=3, option, option)	X = fitted value Y = square root (standardized residual)
autoplot(fit, which=4, option, option)	X = observation number Y = Cook's Distance
autoplot(fit, which=5, option, option)	X = leverage Y = standardized residual
autoplot(fit, which=6, option, option)	X = leverage Y = Cook's Distance
autoplot(fit, which=1:4, option, option)	Note - You can select which plots you want.

3. One Way Analysis of Variance

initialize session

```
setwd("/cloud/project")                                    # Set working directory
getwd()                                                    # Check working directory
options(scipen=999)                                        # Turn off scientific notation
rm(list = ls())                                            # Clear the Decks
```

import excel source data

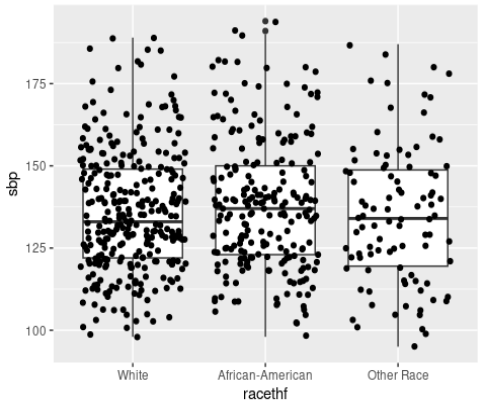
```
library(readxl)
source <- read_excel("hers_640anova.xlsx")
source <- as.data.frame(source)
str(source)

## 'data.frame':    612 obs. of  3 variables:
##  $ raceth : num  3 3 3 3 3 3 3 3 3 ...
##  $ physact: num  1 3 2 5 2 1 1 1 1 ...
##  $ sbp    : num  132 168 105 159 155 126 107 112 166 150 ...
```

3.1 Prepare data: create factors and set reference

```
source$racethf <- factor(source$raceth,                    # Tip: Set explicitly
                         levels=c(1,2,3),
                         labels=c("White", "African-American", "Other Race"))

source$racethf <- relevel(source$racethf, ref="White")     # relevel() with option ref = to set ref
```

3.2 Data Description - Graphical: basic

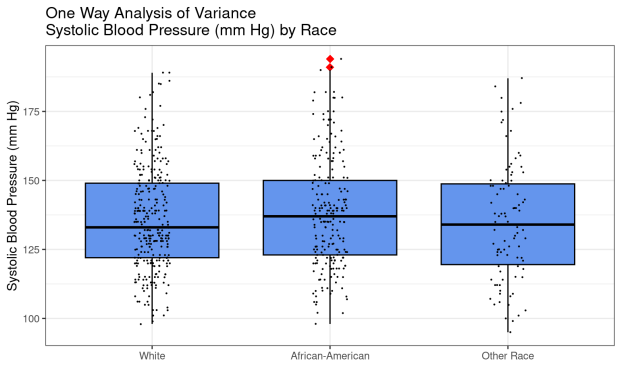
```
library(ggplot2)

# Side-by-side box plot w overlay scatter:  basic
ggplot(data=source) +
  aes(x=racethf) +                                         # x = factor predictor
  aes(y=sbp) +                                             # y = outcome
  geom_boxplot( ) +
  geom_jitter( )
```



Basic plot doesn't look so good. See next page for suggested fixes

3.2 Data Description - Graphical: with added aesthetics

```
library(ggplot2)

# Side-by-side box plot w overlay scatter:  with optional aesthetics
ggplot(data=source) +
  aes(x=racethf) +                              # x = factor predictor
  aes(y=sbp) +                                  # y = outcome

  geom_boxplot(color="black",                   # Tip. boxplot first. color=border of box
               fill= "cornflowerblue",          # fill = color of box fill
               outlier.colour="red",           # user sets outlier color
               outlier.shape=18,                # user sets outlier shape (18=diamond)
               outlier.size=3) +                # user sets outlier size

  geom_jitter(color="black",                    # Overlay plot of scatter
              width=.1,                         # width and height to set extent of jitter
              height=.1,
              size=.1) +                        # size to set size of points

  ggtitle("One Way Analysis of Variance\nSystolic Blood Pressure (mm Hg) by Race") +
  xlab("") +
  ylab("Systolic Blood Pressure (mm Hg)") +
  theme_bw()
```



Better. Also, outliers are now shown in red diamonds

3.2 Data Description, by group - Numerical: Method 1 - basic

```
cat("\nDescriptives by group using summary( )\n")
## Descriptives by group using summary( )

by(source$sbp,source$racethf,summary)

## source$racethf: White
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      98     122     133     136     149     189
## ------------------------------------------------------------
## source$racethf: African-American
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    98.0   123.0   137.0   138.2   150.0   194.0
## ------------------------------------------------------------
## source$racethf: Other Race
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    95.0   119.5   134.0   135.2   148.8   187.0
```

Quick and easy preliminary inspection of the data, even if it's not very nice looking.

3.2 Data Description, by group - Numerical: Method 2 - descr( ) in {summarytools}

```
library(summarytools)

cat("\nDescriptives by group using descr() in {summarytools}\n")
## Descriptives by group using descr() in {summarytools}

with(source,                                                    # with(dataframe,
 stby(data = sbp,                                               # data = yvar
 INDICES =racethf,                                              # INDICES = factor var
 FUN = descr,
 stats=c("n.valid", "pct.valid", "mean", "sd", "min", "max"),   # user chooses statistics to show
 #stats=c("common"),                                            # NOT RUN: another set to show
 transpose=TRUE))

## Descriptive Statistics
## sbp by racethf
## Data Frame: source
## N: 300
##
##                          N.Valid   Pct.Valid     Mean   Std.Dev     Min      Max
## ---------------------- --------- ----------- -------- --------- ------- --------
##                  White    300.00      100.00   136.01     18.55   98.00   189.00
##       African-American    218.00      100.00   138.23     19.99   98.00   194.00
##             Other Race     94.00      100.00   135.18     21.26   95.00   187.00

descr() has the advantage of lots of options (choices of statistics, layout, etc)
```

3.2 Data Description, by group - Numerical: Method 3 - custom

```
library(tidyverse)
library(knitr)                                                  # kable() in {knitr} for nice looking table

mydescriptives <- source %>%
                  group_by(racethf) %>%                         # For each level of factor racethf, obtain:
                  summarise(
                        n = sum(!is.na(sbp)),                         # subgroup sample size (complete obs only)
                        mean=mean(sbp,na.rm=TRUE),                    # mean
                        sd=sd(sbp,na.rm=TRUE),                        # standard deviation
                        se=sd/sqrt((n)),                              # standard error
                        'lower 95% CI' = mean - qt(0.975, n-1)*se,    # lower 95% CI using t-distribution
                        'upper 95% CI' = mean + qt(0.975, n-1)*se,    # upper 95% CI using t-distribution
                         )


cat("\nDescriptives by group using dplyr() and kable() in {knitr}\n")
## Descriptives by group using dplyr() and kable() in {knitr}

kable(mydescriptives, digits=2,                                 # show
      caption="Systolic Blood Pressure (mm Hg), by Race")       # caption = "your title"
```

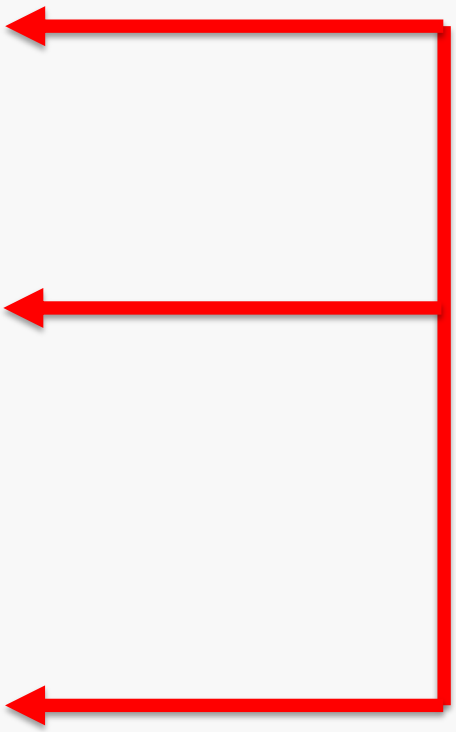
Systolic Blood Pressure (mm Hg), by Race

racethf	n	mean	sd	se	lower 95% CI	upper 95% CI
White	300	136.01	18.55	1.07	133.91	138.12
African-American	218	138.23	19.99	1.35	135.57	140.90
Other Race	94	135.18	21.26	2.19	130.83	139.54

kable( ) produces such pretty tables!

In a ONE WAY ANOVA, the type I, II and II SSQ are the same

```
library(car)


fit <- aov(sbp ~ racethf, data=source)


cat("\nType I ssq\n")
## Type I ssq
anova(fit)
## Analysis of Variance Table
##
## Response: sbp
##            Df Sum Sq Mean Sq F value Pr(>F)
## racethf     2    871  435.50  1.1448  0.319
## Residuals 609 231671  380.41


cat("\n\nType II SSQ\n")
## Type II SSQ
Anova(fit, type="II")
## Anova Table (Type II tests)
##
## Response: sbp
##           Sum Sq  Df F value Pr(>F)
## racethf      871   2  1.1448  0.319
## Residuals 231671 609



cat("\n\nType III SSQ\n")
## Type III SSQ
Anova(fit, type="III")
## Anova Table (Type III tests)
##
## Response: sbp
##              Sum Sq  Df    F value              Pr(>F)
## (Intercept) 5549888   1 14589.1488 <0.0000000000000002 ***
## racethf         871   2     1.1448               0.319
## Residuals    231671 609
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



3.3 Model Estimation and Interpretation: Method 1 - Using aov( )

```
m1_anova <- aov(sbp ~ racethf, data=source)                                    # aov(yvar ~ factorvar,data=DATAFRAME)


cat("\nOne Way Anova using aov() and anova()\n")
## One Way Anova using aov() and anova()

anova(m1_anova)                                                                # anova(MODELOBJECT) to show results
## Analysis of Variance Table
##
## Response: sbp
##            Df Sum Sq Mean Sq F value Pr(>F)
## racethf     2    871  435.50  1.1448  0.319      Overall F test (Null: equality of means) does NOT reject null (p=.32)
## Residuals 609 231671  380.41

cat("\n\nOne Way Anova using aov() and summary()\n")
## One Way Anova using aov() and summary()

summary(m1_anova)                                                              # summary(MODELFIT) to show results
##              Df Sum Sq Mean Sq F value Pr(>F)
## racethf       2    871   435.5   1.145  0.319
## Residuals   609 231671   380.4

anova( ) provides a bit more information than summary( )
```

3.3 Model Estimation and Interpretation: Method 2 - lm( ) and as.factor( )

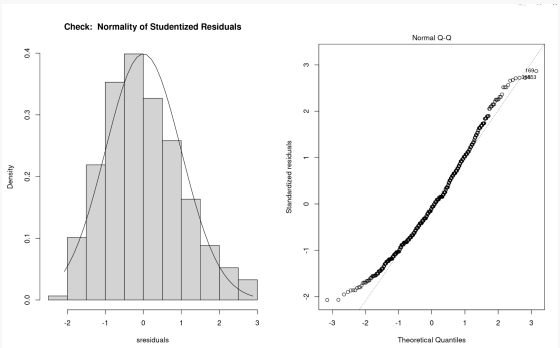
```
m1_regression <- lm(sbp ~ as.factor(raceth), data=source)                      # lm(yvar ~ as.factor(groupvar), data=)

cat("\nOne Way Anova using lm(), as.factor() and anova()\n")
## One Way Anova using lm(), as.factor() and anova()

anova(m1_regression)
## Analysis of Variance Table
##
## Response: sbp
##                    Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(raceth)   2    871  435.50  1.1448  0.319
## Residuals         609 231671  380.41

cat("\n\nOne Way Anova using lm(), as.factor() and summary()\n")
## One Way Anova using lm(), as.factor() and summary()

summary(m1_regression)                                                         # summary(MODELFIT) to show results
## Call:
## lm(formula = sbp ~ as.factor(raceth), data = source)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.234 -14.234  -1.624  12.766  55.766
##
## Coefficients:
##                    Estimate Std. Error t value            Pr(>|t|)
## (Intercept)        136.0133     1.1261 120.786 <0.0000000000000002 ***
## as.factor(raceth)2   2.2206     1.7358   1.279               0.201            Not surprising: no effect of racethf = 2
## as.factor(raceth)3  -0.8325     2.3054  -0.361               0.718                 Similarly, no effect of racethf = 3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.5 on 609 degrees of freedom
## Multiple R-squared:  0.003746,   Adjusted R-squared:  0.0004738               Not surprising that R-squared is tiny!
## F-statistic: 1.145 on 2 and 609 DF,  p-value: 0.319
```

3.3 Model Estimation and Interpretation: Method 3 - lm( ) and indicator vars

```
library(tidyverse)

source <- source %>%
          mutate(African_American = ifelse(racethf=="African-American",1,0)) %>%
          mutate(Other_Race = ifelse(racethf=="Other Race",1,0))

m1_regression2 <- lm(sbp ~ African_American + Other_Race, data=source)    # lm(yvar ~ as.factor(groupvar), data=DATAFRAME)

cat("\nOne Way Anova using lm(), indicator vars() and anova()\n")
## One Way Anova using lm(), indicator vars() and anova()

anova(m1_regression2)
## Analysis of Variance Table
##
## Response: sbp
##                   Df Sum Sq Mean Sq F value Pr(>F)
## African_American   1    821  821.40  2.1592 0.1422              I like using 0/1 indicators; output is more readable
## Other_Race         1     50   49.60  0.1304 0.7182
## Residuals        609 231671  380.41

cat("\n\nOne Way Anova using lm(), indicator vars() and summary()\n")
## One Way Anova using lm(), indicator vars() and summary()

summary(m1_regression2)
## Call:
## lm(formula = sbp ~ African_American + Other_Race, data = source)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.234 -14.234  -1.624  12.766  55.766
##
## Coefficients:
##                  Estimate Std. Error t value            Pr(>|t|)
## (Intercept)      136.0133     1.1261 120.786 <0.0000000000000002 ***
## African_American   2.2206     1.7358   1.279               0.201
## Other_Race        -0.8325     2.3054  -0.361               0.718
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.5 on 609 degrees of freedom
## Multiple R-squared:  0.003746,   Adjusted R-squared:  0.0004738
## F-statistic: 1.145 on 2 and 609 DF,  p-value: 0.319
```

3.4 Regression Diagnostics for ANOVA: normality of residuals

```
library(MASS)

sresiduals <- studres(m1_anova)                    # Null: studentized residuals are Normal(0,1)
par(mfrow = c(1,2))                                # set graph to be 2 panes (1 row, 2 col)
hist(sresiduals,                                   # histogram of sresiduals (plot density not freq)
     freq=FALSE,
     main="Check:  Normality of Studentized Residuals")
xfit <- seq(min(sresiduals),max(sresiduals),length=40)
yfit <- dnorm(xfit)
lines(xfit,yfit)
plot(m1_anova, which = 2)                                # which=2 qqplot (look for straight line)
```



Not great, but sometimes the cure is worse than the problem. Onward

```
par(mfrow = c(1,1))                                # restore graph to be 1 pane (1 row, 1 col)
```

3.4 Regression Diagnostics for ANOVA: test of normality of residuals

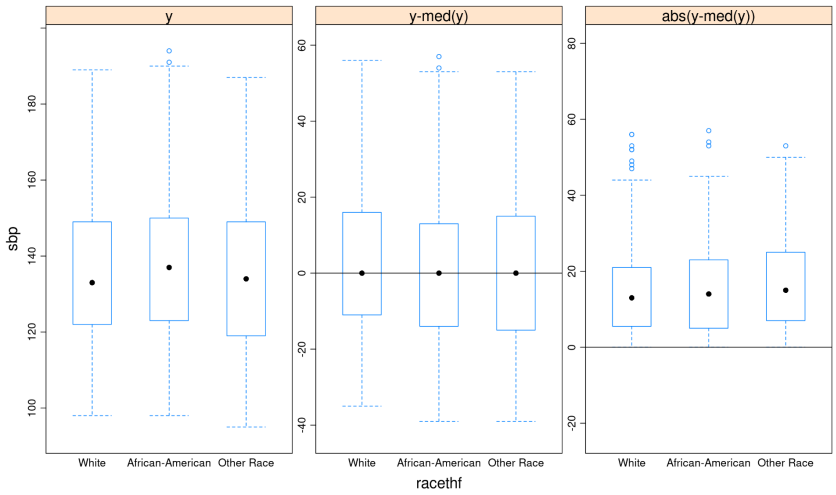
```
shapiro.test(source$fit.resid)
##  Shapiro-Wilk normality test
##
## data:  source$fit.resid
## W = 0.98034, p-value = 0.0000002518          Test of Null: normality of residuals is rejected, possibly due to large n
```

3.4 Regression Diagnostics for ANOVA: test of constant variance of residuals

```
library(HH)                                        # hov() for Brown-Forsyth in package {HH}
library(car)                                       # leveneTest() in package {car}

# null:  constant variance, all is well
bartlett.test(sbp ~ racethf, data=source)          # argument must be model formula
##  Bartlett test of homogeneity of variances
##
## data:  sbp by racethf
## Bartlett's K-squared = 3.1766, df = 2, p-value = 0.2043        Test of Null: constant variance is NOT rejected (good!)

hov(sbp ~ racethf, data=source)                    # argument must be model formula
##  hov: Brown-Forsyth
##
## data:  sbp
## F = 1.4702, df:racethf = 2, df:Residuals = 609, p-value = 0.2307     Again, null of constant variance is NOT rejected
## alternative hypothesis: variances are not identical

leveneTest(sbp ~ racethf, data=source)
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   2  1.4702 0.2307                  leveneTest( ) in {car} is the same as hov( ) in {HH} but a bit more readable
##       609
```

3.4 Regression Diagnostics for ANOVA: Plot of variance of residuals

```
library(HH)                                                    # hovPlot() in package {HH}

hovPlot(sbp ~ racethf, data=source)
```



Plot is consistent with results of tests of common variance (null not rejected)

3.5 Post-hoc Tests and Estimation: Pairwise t-tests

```
cat("\nOne Way Anova: Pairwise t-tests - No adjustment for multiple comparisons\n")
## One Way Anova: Pairwise t-tests - No adjustment for multiple comparisons

pairwise.t.test(source$sbp,source$racethf,p.adjust.method = "none")
##  Pairwise comparisons using t tests with pooled SD
##
## data:  source$sbp and source$racethf
##
##                 White African-American
## African-American 0.20  -
## Other Race       0.72  0.21                          No statistically significant pairwise comparison of group means
##
## P value adjustment method: none

cat("\n\nOne Way Anova: Pairwise t-tests - Bonferroni adjustment\n"
## One Way Anova: Pairwise t-tests - Bonferroni adjustment

pairwise.t.test(source$sbp,source$racethf,p.adjust.method = "bonferroni")
##  Pairwise comparisons using t tests with pooled SD
##
## data:  source$sbp and source$racethf
##
##                 White African-American
## African-American 0.60  -
## Other Race       1.00  0.62                           Ajustment for multiple comparisons yields "more null" p-values
##
## P value adjustment method: bonferroni
```

3.5 Post-hoc Tests and Estimation: Tukey pairwise comparison of means

```
# TIP. Providing the formula instead of the model object produces output that is easier to read.
TukeyHSD(aov(sbp~racethf, data=source))
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = sbp ~ racethf, data = source)
##
## $racethf
##                                   diff       lwr      upr     p adj
## African-American-White       2.2206116 -1.857604 6.298827 0.4074184
## Other Race-White            -0.8324823 -6.248970 4.584005 0.9306610
## Other Race-African-American -3.0530939 -8.707400 2.601212 0.4135187 Nice.  We also get 95% CI's of differences in means
```

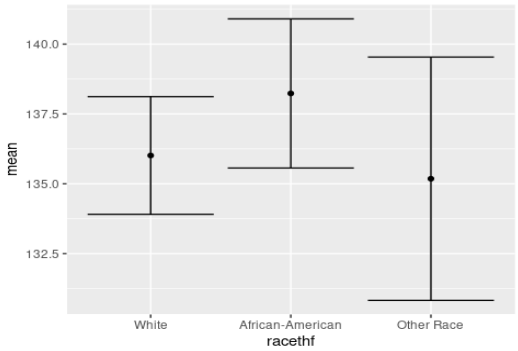
3.5 Post-hoc Tests and Estimation: general linear model contrasts

```
library(multcomp)                                                    # glht( ) in package {multcomp}

glht(m1_anova, linfct = mcp(racethf = "Tukey"))
##   General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Linear Hypotheses:
##                                    Estimate
## African-American - White == 0        2.2206
## Other Race - White == 0             -0.8325
## Other Race - African-American == 0  -3.0531       glht( ) produces Tukey pairwise comparison of means (matches above)
```

```
summary(glht(m1_anova, linfct = mcp(racethf = "Tukey")))
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = sbp ~ racethf, data = source)
##
## Linear Hypotheses:
##                                    Estimate Std. Error t value Pr(>|t|)
## African-American - White == 0        2.2206     1.7358   1.279    0.403
## Other Race - White == 0             -0.8325     2.3054  -0.361    0.930
## Other Race - African-American == 0  -3.0531     2.4066  -1.269    0.409
## (Adjusted p values reported -- single-step method)                    summary(glht( )) gives a bit more information
```

3.6 Post-hoc Visualization: group means with 95% - basic

```
library(ggplot2)
library(tidyverse)

# get descriptives for plotting
plotdata <- source %>%                    # create df for plotting that contains within group statistics
   group_by(racethf) %>%                        # for each level of racethf
   summarise(                                   # obtain the following summaries:
        n = sum(!is.na(sbp)),                            # sample size (complete observations only)
        mean = mean(sbp, na.rm=TRUE),                    # mean (remove missings)
        sd = sd(sbp, na.rm=TRUE),                        # standard deviation (remove missings)
        se = sd/sqrt(n),                                 # standard error
        tcoef = qt(0.975, n -1),                         # Student-t multiplier for 95% CI
        lower_CI = mean - tcoef*se,                      # lower CI limit
        upper_CI = mean + tcoef*se)                      # upper CI limit


plotdata                                         # show
## # A tibble: 3 × 8
##   racethf              n  mean    sd    se tcoef lower_CI upper_CI
##   <fct>            <int> <dbl> <dbl> <dbl> <dbl>    <dbl>    <dbl>
## 1 White              300  136.  18.6  1.07  1.97     134.     138.
## 2 African-American   218  138.  20.0  1.35  1.97     136.     141.
## 3 Other Race          94  135.  21.3  2.19  1.99     131.     140.

# Basic plot - required layers only & no frills
ggplot(data=plotdata) +
   aes(x=racethf) +
   aes(y=mean) +
   geom_errorbar(aes(ymin=lower_CI, ymax=upper_CI))  +
   geom_point()
```

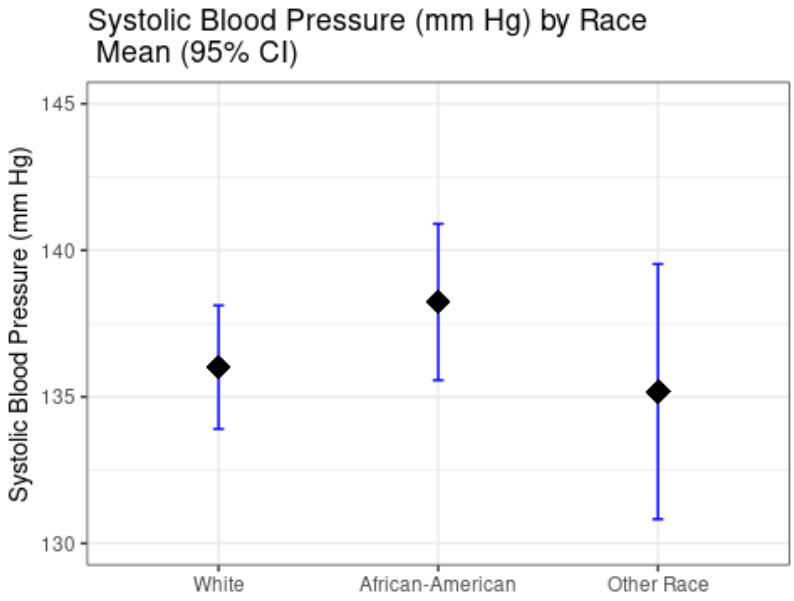


Okay looking but this plot could use some aesthetics!

3.6 Post-hoc Visualization: group means with 95% - with aesthetics

```
library(ggplot2)
library(tidyverse)

# get descriptives for plotting
plotdata <- source %>%                       # create df for plotting that contains within group statistics
   group_by(racethf) %>%                           # for each group defined by racethf
   summarise(                                      # obtain the following summaries:
        n = sum(!is.na(sbp)),                              # sample size (complete observations only)
        mean = mean(sbp, na.rm=TRUE),                      # mean (remove missings)
        sd = sd(sbp, na.rm=TRUE),                          # standard deviation (remove missings)
        se = sd/sqrt(n),                                   # standard error
        tcoef = qt(0.975, n -1),                           # Student-t multiplier for 95% CI
        lower_CI = mean - tcoef*se,                        # lower CI limit
        upper_CI = mean + tcoef*se)                        # upper CI limit


#plotdata                                    # NOT RUN AGAIN:  show
```

```
# Plot with aesthetics added
ggplot(data=plotdata) +
   aes(x=racethf) +
   aes(y=mean) +

   geom_errorbar(aes(ymin=lower_CI, ymax=upper_CI),
                 color="blue", width=.05)  +

   geom_point(color="black", shape=18, size=5) +

  scale_y_continuous(limits=c(130, 145),
                     breaks=c(130,135,140,145)) +

  ggtitle("Systolic Blood Pressure (mm Hg) by Race\n Mean (95% CI)") +
  xlab("") +
  ylab("Systolic Blood Pressure (mm Hg)") +
  theme_bw()
```



This looks better!

4. Two Way Factorial Analysis of Variance

In this illustration of a two-way factorial anova, we will investigate the statistical significance of differences in the mean value of `sbp` due to: 1) a main effect of `factor I = racethf`; 2) a main effect of `factor II = activityc`, a new variable that is physact collapsed to 3 levels; and 3) the `interaction racethf x activityc`.

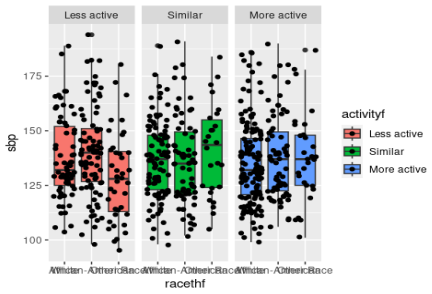
A challenge of performing a two-factorial analysis of variance pertains to which partial F-tests you want to perform and in what order. Because the interpretation of a main effect of a factor (I or II) will be different depending on whether or not there is an interaction of the two factors (I x II), a reasonable approach is to begin the analysis with a test of the null hypothesis of zero interaction.

4.1 Prepare data: create factors and set reference levels

```
library(tidyverse)
library(summarytools)

#  Factor I:  raceth at 3 levels
source$racethf <- factor(source$raceth,
                         levels=c(1,2,3),
                         labels=c("White", "African-American", "Other Race"))
source$racethf <- relevel(source$racethf, ref="White")


#  Factor II:  For illustration, create activityc = new summary measure of physical activity at 3 levels
source <- source %>%
   mutate(activityc = case_when(
               physact %in% 1:2 ~ "1",
                     physact==3 ~ "2",
               physact %in% 4:5 ~ "3")) %>%

    mutate(activityf = factor(activityc,
                              levels = c("1","2", "3"),
                              labels = c("Less active", "Similar", "More active")))

source$activityf <- relevel(source$activityf, ref="Less active")

cat("\nCHECK:  Creation of activityf")
ctable(x=source$physact, y=source$activityf, prop="n")

## CHECK:  Creation of activityf
## Cross-Tabulation
## physact * activityf
## Data Frame: source
##
## --------- ----------- ------------- --------- ------------- -------
##             activityf   Less active   Similar   More active   Total
##   physact
##         1                        65         0             0      65
##         2                       127         0             0     127
##         3                         0       192             0     192
##         4                         0         0           165     165
##         5                         0         0            63      63
##     Total                       192       192           228     612
## --------- ----------- ------------- --------- ------------- -------
```

4.2 Data Description - Graphical: basic

```
library(tidyverse)
library(ggplot2)

ggplot(data=source) +
  aes(x=racethf) +                              # x = factor predictor
  aes(y=sbp) +                                  # y = outcome
  aes(fill=activityf) +                         # fill = stratification variable
  geom_boxplot( ) +                             # Tip.  Plot boxplot first
  geom_jitter( ) +                              # Tip.  Overlay jitter plot on top
  facet_grid(.~activityf)                       #  panels in 1 row
```

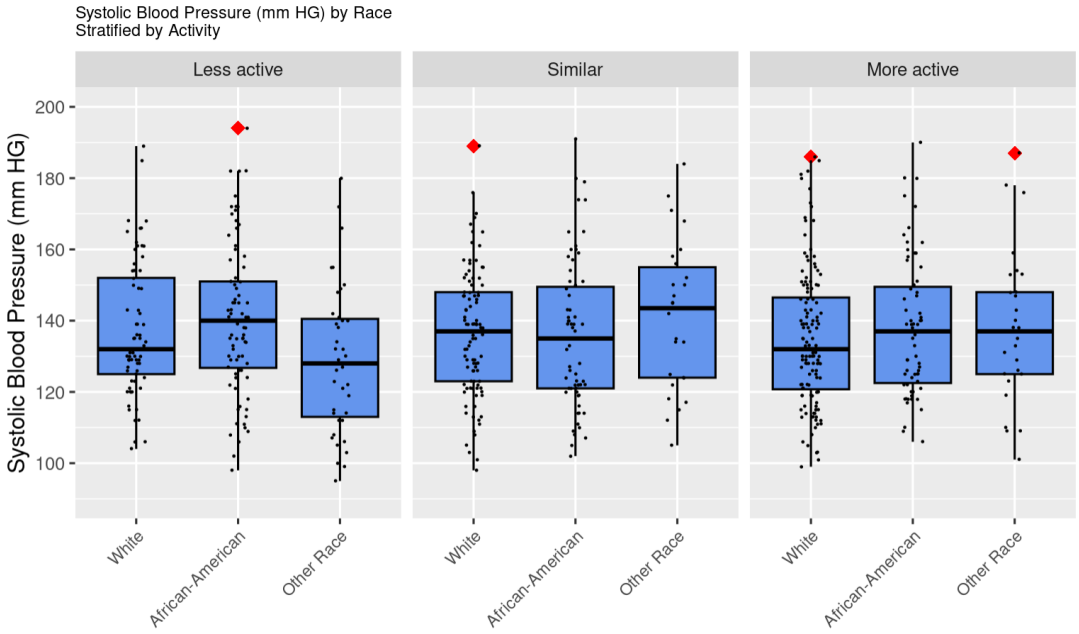


Basic graph does not look good. Needs fixing!

```
  #facet_grid(activityf ~.)                    #  NOT RUN:  how to set panels in 1 column
```

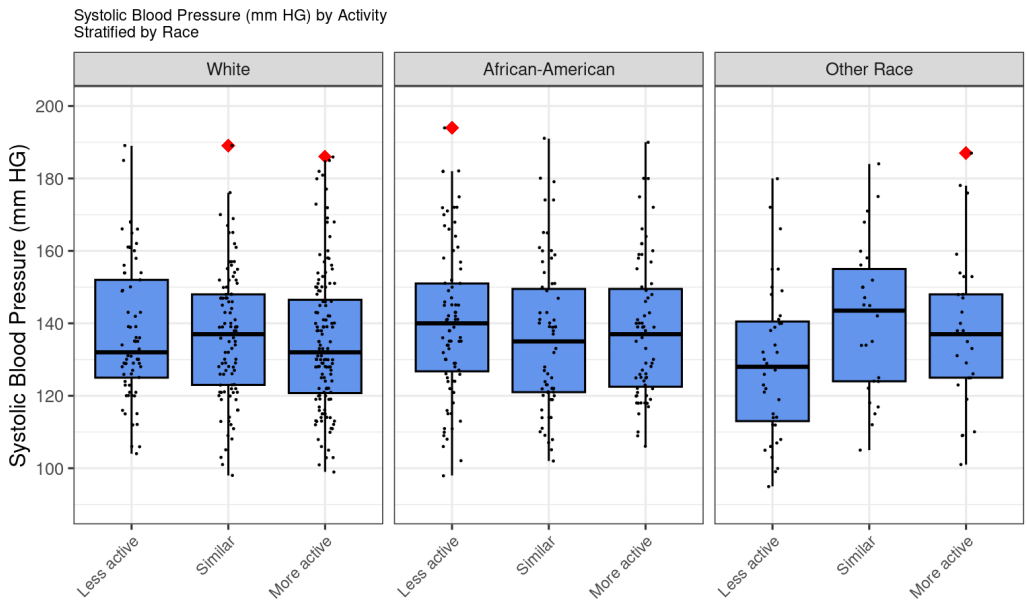
4.2 Data Description - Graphical: with aesthetics, X=racethf, strata=activityf

```
library(tidyverse)
library(ggplot2)

# get min and max of Y=sbp for setting Y-axis tick marks
min(source$sbp)
## [1] 95

max(source$sbp)
## [1] 194

# Y=sbp,  X=racethf,  Strata=activityf
ggplot(data=source) +
  aes(x=racethf) +                              # x = factor predictor
  aes(y=sbp) +                                  # y = outcome
  aes(fill=activityf) +                         # fill = stratification variable
  geom_boxplot(color="black",
       fill= "cornflowerblue",
       outlier.colour="red",
       outlier.shape=18,
       outlier.size=3) +
  geom_jitter(color="black",
       width=.1,
       height=.1,
       size=.1) +
  facet_grid(.~activityf) +
  scale_y_continuous(limits=c(90, 200),
                     breaks=c(100, 120, 140, 160, 180,200)) +
   xlab("") +
   ylab("Systolic Blood Pressure (mm HG)") +
   ggtitle("Systolic Blood Pressure (mm HG) by Race\nStratified by Activity") +
   theme(plot.title=element_text(size=8),
          axis.text.x = element_text(size=8, angle=45, hjust=1),
          legend.position = "none")
```



4.2 Data Description - Graphical: with aesthetics, X=activityf, strata=racethf

```
library(tidyverse)
library(ggplot2)

# get min and max of Y=sbp for setting Y-axis tick marks
min(source$sbp)
## [1] 95

max(source$sbp)
## [1] 194

# Y=sbp,  X=activityf,  Strata=racef
ggplot(data=source) +
  aes(x=activityf) +                          # x = factor predictor
  aes(y=sbp) +                                # y = outcome
  aes(fill=racethf) +                         # fill = stratification variable
  geom_boxplot(color="black",
       fill= "cornflowerblue",
       outlier.colour="red",
       outlier.shape=18,
       outlier.size=3) +
  geom_jitter(color="black",
       width=.1,
       height=.1,
       size=.1) +
  facet_grid(.~racethf) +
  scale_y_continuous(limits=c(90, 200),
                     breaks=c(100, 120, 140, 160, 180,200)) +
   xlab("") +
   ylab("Systolic Blood Pressure (mm HG)") +
   ggtitle("Systolic Blood Pressure (mm HG) by Activity\nStratified by Race") +
   theme_bw() +
   theme(plot.title=element_text(size=8),
          axis.text.x = element_text(size=8, angle=45, hjust=1),
          legend.position = "none")
```



4.2 Data Description, by group - Numerical: Method 1

```
library(FSA)                                                    #  Summarize() in package {FSA}

cat("\nTwo Way Anova: descriptives by group using Summarize( ) in {FSA}\n")
## Two Way Anova: descriptives by group using Summarize( ) in {FSA}

Summarize(sbp ~ racethf + activityf, digits=2, data=source)                  Summarize( ) in {FSA} is quick and easy
##            racethf   activityf   n   mean    sd min     Q1 median    Q3 max
## 1            White Less active  69 137.30 18.76 104 125.00  132.0 152.0 189
## 2 African-American Less active  84 140.18 20.48  98 126.75  140.0 151.0 194
## 3       Other Race Less active  39 128.92 20.67  95 113.00  128.0 140.5 180
## 4            White     Similar  99 136.53 17.62  98 123.00  137.0 148.0 189
## 5 African-American     Similar  67 136.10 20.28 102 121.00  135.0 149.5 191
## 6       Other Race     Similar  26 141.08 21.08 105 124.00  143.5 155.0 184
## 7            White More active 132 134.95 19.19  99 120.75  132.0 146.5 186
## 8 African-American More active  67 137.93 19.13 106 122.50  137.0 149.5 190
## 9       Other Race More active  29 138.31 20.67 101 125.00  137.0 148.0 187
```

4.2 Data Description, by group - Numerical: Method 2 - custom

```
library(tidyverse)
library(knitr)

mydescriptives2 <- source %>%
                  group_by(racethf,activityf) %>%
                  summarise(
                        n=n(),
                        mean=mean(sbp,na.rm=TRUE),
                        sd=sd(sbp,na.rm=TRUE),
                        se=sd/sqrt((n)),
                        'lower 95% CI' = mean - qt(0.975, n-1)*se,
                        'upper 95% CI' = mean + qt(0.975, n-1)*se)

kable(mydescriptives2, digits=2,
      caption="Systolic Blood Pressure (mm Hg), by Race and Activity")
```

Systolic Blood Pressure (mm Hg), by Race and Activity

racethf	activityf	n	mean	sd	se	lower 95% CI	upper 95% CI
White	Less active	69	137.30	18.76	2.26	132.80	141.81
White	Similar	99	136.53	17.62	1.77	133.01	140.04
White	More active	132	134.95	19.19	1.67	131.65	138.26
African-American	Less active	84	140.18	20.48	2.23	135.73	144.62
African-American	Similar	67	136.10	20.28	2.48	131.16	141.05
African-American	More active	67	137.93	19.13	2.34	133.26	142.59
Other Race	Less active	39	128.92	20.67	3.31	122.22	135.63
Other Race	Similar	26	141.08	21.08	4.13	132.56	149.59
Other Race	More active	29	138.31	20.67	3.84	130.45	146.17

This looks better.

4.3 Model Estimation and Interpretation: Method 1 - Using aov( )

```
library(car)                                                    # Anova() in package {car}

# Tip.  Order predictors for interpretability of Type I SSQ:
# yvar ~ main_effect + main_effect + interaction
m2_anova <- aov(sbp ~ racethf + activityf + racethf:activityf, data=source)


cat("\nTwo Way Factorial Anova:  Type I, II, III SSQ are NOT identical\n")
cat("\nType I ssq\n")
anova(m2_anova)

## Two Way Factorial Anova:  Type I, II, III SSQ are NOT identical
## Type I ssq
## Analysis of Variance Table
##
## Response: sbp
##                    Df Sum Sq Mean Sq F value  Pr(>F)
## racethf             2    871  435.50  1.1516 0.31683
## activityf           2     41   20.73  0.0548 0.94666
## racethf:activityf   4   3590  897.60  2.3735 0.05109 .
## Residuals         603 228039  378.17
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

cat("\n\nType II SSQ\n")
Anova(m2_anova, type="II")

## Type II SSQ
## Anova Table (Type II tests)
##
## Response: sbp
##                   Sum Sq  Df F value  Pr(>F)
## racethf              846   2  1.1185 0.32745
## activityf             41   2  0.0548 0.94666
## racethf:activityf   3590   4  2.3735 0.05109 .
## Residuals         228039 603
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cat("\n\nType III SSQ\n")
Anova(m2_anova, type="III")

## Type III SSQ
## Anova Table (Type III tests)
##
## Response: sbp
##                    Sum Sq  Df   F value               Pr(>F)
## (Intercept)       1300821   1 3439.7409 < 0.0000000000000002 ***
## racethf              3396   2    4.4893              0.01161 *
## activityf             289   2    0.3820              0.68266
## racethf:activityf    3590   4    2.3735              0.05109 .
## Residuals          228039 603
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.3 Model Estimation and Interpretation: Method 2 - lm() and indicator vars

```
library(tidyverse)

source <- source %>%
        # Indicators for main effects
        mutate(I_racea = ifelse(racethf=="African-American",1,0)) %>%     # (3-1) 0/1's for racethf at 3 levels
        mutate(I_raceo = ifelse(racethf=="Other Race",1,0)) %>%
        mutate(I_actives = ifelse(activityf=="Similar",1,0)) %>%          # (3-1) 0/1's for activityf at 3 levels
        mutate(I_activem = ifelse(activityf=="More active",1,0)) %>%
        # Indicators for interactions
        mutate(raceaXactives = I_racea*I_actives) %>%                     # interactions
        mutate(raceaXactivem = I_racea*I_activem) %>%
        mutate(raceoXactives = I_raceo*I_actives) %>%
        mutate(raceoXactivem = I_raceo*I_activem)

m2_regression <- lm(data=source,
                 sbp ~ I_racea + I_raceo + I_actives + I_activem +
                       raceaXactives + raceaXactivem + raceoXactives + raceoXactivem)

cat("\nTwo Way Anova using lm(), indicator vars, and anova()\n")
anova(m2_regression)


## Two Way Anova using lm(), indicator vars, and anova()
## Analysis of Variance Table
##
## Response: sbp
##                Df Sum Sq Mean Sq F value  Pr(>F)
## I_racea         1    821  821.40  2.1720 0.14106
## I_raceo         1     50   49.60  0.1312 0.71736
## I_actives       1     29   28.57  0.0755 0.78352
## I_activem       1     13   12.89  0.0341 0.85358
## raceaXactives   1    887  886.75  2.3448 0.12622
## raceaXactivem   1    277  277.01  0.7325 0.39242
## raceoXactives   1    751  750.61  1.9848 0.15940      A plug for using explictly defined 0/1 indicators:
## raceoXactivem   1   1676 1676.06  4.4320 0.03568 *    0/1 indicators lets us see the marginally significant interaction
## Residuals     603 228039  378.17
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cat("\n\nTwo Way Anova using lm(), indicator vars, and summary()\n")
summary(m2_regression)

## Two Way Anova using lm(), indicator vars, and summary()
## Call:
## lm(formula = sbp ~ I_racea + I_raceo + I_actives + I_activem +
##     raceaXactives + raceaXactivem + raceoXactives + raceoXactivem,
##     data = source)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.179 -14.360  -1.418  11.885  54.896
##
## Coefficients:
##               Estimate Std. Error t value            Pr(>|t|)
## (Intercept)   137.3043     2.3411  58.649 <0.0000000000000002 ***
## I_racea         2.8742     3.1596   0.910              0.3633
## I_raceo        -8.3813     3.8958  -2.151              0.0318 *
## I_actives      -0.7791     3.0497  -0.255              0.7985
## I_activem      -2.3498     2.8889  -0.813              0.4163
## raceaXactives  -3.2950     4.4099  -0.747              0.4552
## raceaXactivem   0.0966     4.3003   0.022              0.9821
## raceoXactives  12.9329     5.7916   2.233              0.0259 *          Odd. This does NOT match what anova( ) shows
## raceoXactivem  11.7371     5.5752   2.105              0.0357 *                  This DOES match what anova( ) shows
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.45 on 603 degrees of freedom
## Multiple R-squared:  0.01936,    Adjusted R-squared:  0.006354
## F-statistic: 1.488 on 8 and 603 DF,  p-value: 0.1581
```

4.4 Post-hoc Tests and Estimation

```
library(tidyverse)

# First
# Test of Interaction
# Partial F-test of Null: Controlling for main effects, no interaction/effect modification
full1 <- aov(sbp ~ racethf + activityf + racethf:activityf, data=source)
reduced1 <- aov(sbp ~ racethf + activityf, data=source)


cat("\nTwo Way Factorial ANOVA\n")
cat("F-Test of Null:  No interaction\n")
anova(full1, reduced1)

##
## Two Way Factorial ANOVA
## F-Test of Null:  No interaction

## Analysis of Variance Table
##
## Model 1: sbp ~ racethf + activityf + racethf:activityf
## Model 2: sbp ~ racethf + activityf
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    603 228039
## 2    607 231629 -4   -3590.4 2.3735 0.05109 .     Controlling for main effects, interaction is marginally significant
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Second
# Partial F-Test of Null:  No main effect of race in model with ZERO interaction
full2 <- aov(sbp ~ activityf + racethf, data=source)
reduced2 <- aov(sbp ~ activityf, data=source)


cat("\nTwo Way Factorial ANOVA\n")
cat("F-Test of Null:  No Main Effect Race controlling for Activity (assuming NO interaction)\n")
anova(full2, reduced2)

##
## Two Way Factorial ANOVA
## F-Test of Null:  No Main Effect Race controlling for Activity (assuming NO interaction)
## Analysis of Variance Table
##
## Model 1: sbp ~ activityf + racethf
## Model 2: sbp ~ activityf
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    607 231629
## 2    609 232475 -2   -845.96 1.1084 0.3307            Controlling for activity, race is NOT significant

# Second
# Partial F-Test of Null:  No main effect of activity in model with ZERO interaction
full3 <- aov(sbp ~ racethf + activityf, data=source)
reduced3 <- aov(sbp ~ racethf, data=source)


cat("\nTwo Way Factorial ANOVA\n")
cat("F-Test of Null:  No Main Effect Activity controlling for Race (assuming NO interaction)\n")
anova(full3, reduced3)

## Two Way Factorial ANOVA
## F-Test of Null:  No Main Effect Activity controlling for Race (assuming NO interaction)
## Analysis of Variance Table
##
## Model 1: sbp ~ racethf + activityf
## Model 2: sbp ~ racethf
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    607 231629
## 2    609 231671 -2    -41.46 0.0543 0.9471            Controlling for race, activity is NOT significant
```

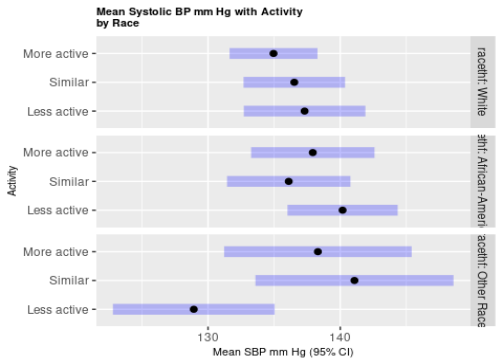
4.4 Post-Hoc Estimation: emmeans() in {emmeans}

```
library(emmeans)

cat("\nTwo Way Anova:  Means by Activity for Strata = Race\n")
## Two Way Anova:  Means by Activity for Strata = Race

emm1 = emmeans::emmeans(m2_anova, specs = "activityf", by="racethf")
emm1
## racethf = White:
##  activityf   emmean   SE  df lower.CL upper.CL
##  Less active    137 2.34 603      133      142
##  Similar        137 1.95 603      133      140
##  More active    135 1.69 603      132      138
##
## racethf = African-American:
##  activityf   emmean   SE  df lower.CL upper.CL
##  Less active    140 2.12 603      136      144
##  Similar        136 2.38 603      131      141
##  More active    138 2.38 603      133      143
##
## racethf = Other Race:
##  activityf   emmean   SE  df lower.CL upper.CL
##  Less active    129 3.11 603      123      135
##  Similar        141 3.81 603      134      149
##  More active    138 3.61 603      131      145
##
## Confidence level used: 0.95                                    Convenient layout
```

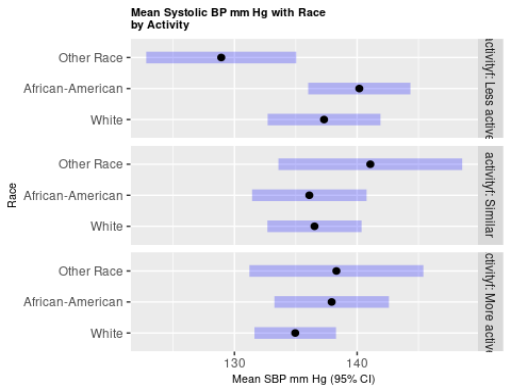
```
cat("\n\nTwo Way Anova:  Means by Race for Strata = Activity\n")
emm2 = emmeans::emmeans(m2_anova, specs = "racethf", by="activityf")

## Two Way Anova:  Means by Race for Strata = Activity

emm2
## activityf = Less active:
##  racethf          emmean   SE  df lower.CL upper.CL
##  White               137 2.34 603      133      142
##  African-American    140 2.12 603      136      144
##  Other Race          129 3.11 603      123      135
##
## activityf = Similar:
##  racethf          emmean   SE  df lower.CL upper.CL
##  White               137 1.95 603      133      140
##  African-American    136 2.38 603      131      141
##  Other Race          141 3.81 603      134      149
##
## activityf = More active:
##  racethf          emmean   SE  df lower.CL upper.CL
##  White               135 1.69 603      132      138
##  African-American    138 2.38 603      133      143
##  Other Race          138 3.61 603      131      145
##
## Confidence level used: 0.95
```

4.4 Post-Hoc Estimation: Plot of means and 95% CI

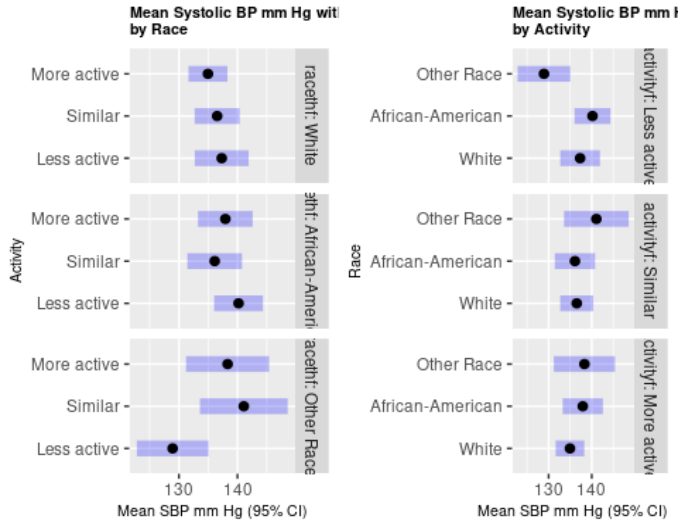
```
library(emmeans)
library(ggplot2)
library(gridExtra)

# Must use lm( ) object
mymodel <- lm(sbp ~ racethf + activityf + racethf:activityf, data = source)

# y=mean sbp  x=activity   strata=race
strata_race <- emmeans (mymodel,  ~ activityf | racethf)
left <- plot(strata_race) +
        labs(x = "Mean SBP mm Hg (95% CI)",
             y = "Activity") +
        ggtitle("Mean Systolic BP mm Hg with Activity \nby Race") +
        theme(axis.title = element_text(size = 8),
             plot.title = element_text(size = 8, face = "bold"))

# y=mean sbp  x=race   strata=activity
strata_activity <- emmeans (mymodel,  ~ racethf | activityf)
right <- plot(strata_activity) +
        labs(x = "Mean SBP mm Hg (95% CI)",
             y = "Race") +
        ggtitle("Mean Systolic BP mm Hg with Race \nby Activity") +
        theme(axis.title = element_text(size = 8),
             plot.title = element_text(size = 8, face = "bold"))

left
```




```
right
```



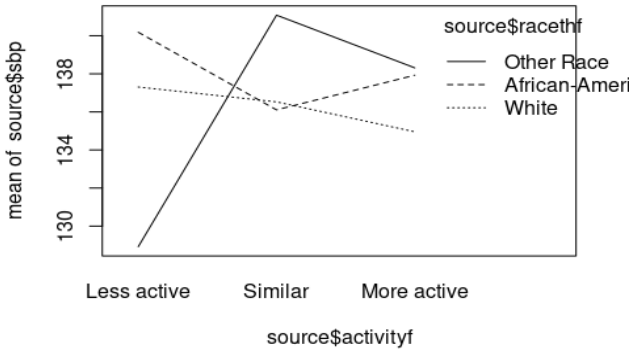

```
gridExtra::grid.arrange(left, right, ncol=2)
```



Combining the 2 graphs into a single graph not a good idea

4.5 Interaction Plot: basic

```
# interaction.plot(GROUPVARIABLE, STRATAVARIABLE, YVARIABLE)
interaction.plot(source$activityf,source$racethf,source$sbp)
```

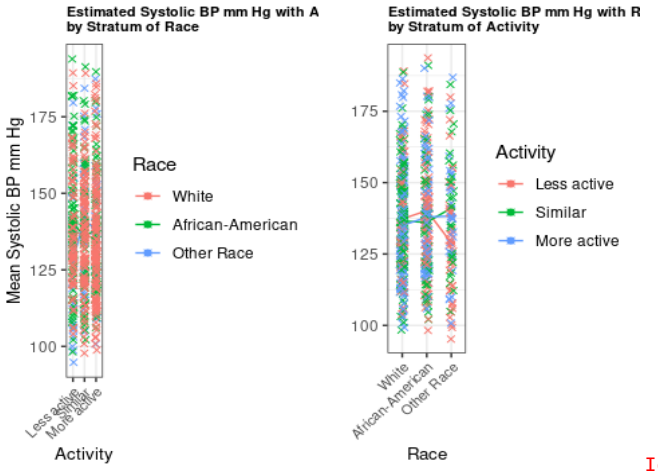


Useful for exploring data but not fancy.

4.5 Interaction Plot: emmip( ) in {emmeans}

```
library(emmeans)
library(ggplot2)
library(gridExtra)

# Must use lm( ) object
mymodel <- lm(sbp ~ racethf + activityf + racethf:activityf, data = source)

# y=sbp  groupvar=activityf    stratumvar=racethf
# emmip(FITOBJECT, STRATIFYVAR ~ XVAR)
left2 <- emmip(mymodel, racethf ~ activityf, style = "factor") +
         geom_jitter(data=source,
                     aes(x = activityf,
                         y = sbp,
                         colour = racethf),
                      pch = 4, width = 0.1) +

       labs(y = "Mean Systolic BP mm Hg",
            x="Activity",
            colour = "Race") +

       ggtitle("Estimated Systolic BP mm Hg with Activity \nby Stratum of Race") +
       theme_bw() +
       theme(axis.title = element_text(size = 10),
             axis.text.x = element_text(size=8, angle=45, hjust=1),
             plot.title = element_text(size = 8, face = "bold"))


# y=sbp  groupvar=racethf    stratumvar=activityf
# emmip(FITOBJECT, STRATIFYVAR ~ XVAR)
right2 <- emmip(mymodel, activityf ~ racethf, style = "factor") +
         geom_jitter(data=source,
                     aes(x = racethf,
                         y = sbp,
                         colour = activityf),
                      pch = 4, width = 0.1) +

       labs(y = "",
            x="Race",
            colour = "Activity") +
```

```
      ggtitle("Estimated Systolic BP mm Hg with Race\nby Stratum of Activity") +
      theme_bw() +
      theme(axis.title = element_text(size = 10),
            axis.text.x = element_text(size=8, angle=45, hjust=1),
            plot.title = element_text(size = 8, face = "bold"))

gridExtra::grid.arrange(left2, right2, ncol=2)
```



Ick! This is an example where an overlay jitter is NOT recommended!

4.5 Interaction Plot: ggplot with aesthetics

```
library(tidyverse)
library(ggplot2)

# get descriptives for plotting
plotdata2 <- source %>%
   group_by(racethf, activityf) %>%
   summarise(
        n = sum(!is.na(sbp)),
        mean = mean(sbp, na.rm=TRUE),
        sd = sd(sbp, na.rm=TRUE),
        se = sd/sqrt(n),
        tcoef = qt(0.975, n -1),
        lower_CI = mean - tcoef*se,
        upper_CI = mean + tcoef*se)

#show
plotdata2
## # A tibble: 9 × 9
## # Groups:   racethf [3]
##   racethf          activityf       n  mean    sd    se tcoef lower_CI upper_CI
##   <fct>            <fct>       <int> <dbl> <dbl> <dbl> <dbl>    <dbl>    <dbl>
## 1 White            Less active    69  137.  18.8  2.26  2.00     133.     142.
## 2 White            Similar        99  137.  17.6  1.77  1.98     133.     140.
## 3 White            More active   132  135.  19.2  1.67  1.98     132.     138.
## 4 African-American Less active    84  140.  20.5  2.23  1.99     136.     145.
## 5 African-American Similar        67  136.  20.3  2.48  2.00     131.     141.
## 6 African-American More active    67  138.  19.1  2.34  2.00     133.     143.
## 7 Other Race       Less active    39  129.  20.7  3.31  2.02     122.     136.
## 8 Other Race       Similar        26  141.  21.1  4.13  2.06     133.     150.
## 9 Other Race       More active    29  138.  20.7  3.84  2.05     130.     146.
```

```
# Y=sbp,  X=activityf,  Strata=racethf
ggplot(data=plotdata2) +
  aes(x=activityf) +                          # x = factor predictor, mean only
  aes(y=mean) +                               # y = outcome
  aes(color=racethf) +                        # fill = stratification variable

  geom_line(aes(group=racethf)) +              # separate line plots by strata
  geom_point() +
  #geom_errorbar(aes(ymin = lower_CI, ymax = upper_CI, width=0.1)) +   # NOT RUN (messy)

  scale_y_continuous(limits=c(125, 145),
                     breaks=c(125, 130, 135, 140, 145)) +

  labs(title = "Systolic Blood Pressure (mm Hg) with Activity",
      subtitle = "Mean (95% CI)",
       x = "Activity Level Compared to Other Women of Same Age",
       y = "mm Hg",
      color="Race") +

  #theme_bw() +                                # NOT run:  remove hashtag to execute (clears gray)
  theme(legend.title=element_blank())         # legend title is blank
```