Unit 6 – Analysis of Variance
Practice Problems (2 of 2)

Due: Wednesday Thursday November 14, 2024
Last date to submit for credit (-20 points): Thursday November 21, 2024

Before you begin. Download from the course website
lbw.xlsx

(Source: Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013) Applied Logistic Regression: Third Edition. These data are copyrighted by John Wiley & Sons Inc. and must be acknowledged and used accordingly. Data were collected at Baystate Medical Center, Springfield, Massachusetts during 1986.)

Low birth weight is an outcome of concern because of its links to infant mortality and birth defects. A woman's behavior during pregnancy (including diet, smoking habits, and receiving prenatal care) can greatly alter the chances of carrying the baby to term and, consequently, of delivering a baby of normal birth weight. The goal of this study was to identify risk factors associated with giving birth to a low birth weight baby (weighing less than 2500 grams). Data were collected on 189 women, 59 of whom had low birth weight babies and 130 of which had normal birth weight babies.

In this homework, we will use three variables to gain practice in performing a two-way analysis of variance: lbw.xlsx has 189 observations on 3 variables.

Data dictionary/Codebook

Position	Variable	Label	Type	Codings
1	id	Identification code		Range: 4 - 226
2	race	Race	numeric	1 = white 2 = african american 3 = other
3	ftv	Number of visits to physician during 1st trimester	numeric	Range: 0 - 6
4	btw	Birthweight (grams)	numeric	Range: 709 - 4990

Outcome Variable
Y = btw

Factor I
racef, coded: 1, 2 or 3
Note: you will create this from race in exercise #2

Factor II
no_trimester1, coded: 0, 1
Note: you will create this from ftv in exercise #2

#1.

State the analysis of variance model using notation μ, α_i, β_j, $(\alpha\beta)_{ij}$ and σ^2 as appropriate. Define all terms and constraints on the parameters.

#2.

By any means you like, create the following three new variables

(1) **racef** = factor version of race

(2) **no_trimester1** that is a 0/1 indicator of "no visits in the first trimester and defined as follows:

no_trimester1 = 1 if **ftv**=0
0 for all other values of **ftv**

(3) **no_trimester1f** = factor version of **no_trimester1**

#3.

By any means you like, produce descriptive statistics of **Y=bwt**, separately for groups defined by **racef** and **no_trimester1f**.

#4.

Fit the two-way analysis of variance. Show the analysis of variance table.

#5.

This time, perform the two way analysis of variance as a regression. Show.

#6.

By any means you like, perform a partial F-test of the null hypothesis that, controlling for **racef** and **no_trimester1f**, the extra predictive significance of the interaction of **racef** and **no_trimester1f** is zero.

#7.

Obtain the predicted means of bwt for each group defined by **racef** and **no_trimester1f**, in two ways: (1) from the analysis of variance; and (2) from the regression. Verify that they are identical.

Supplement - Learn R (nothing to turn in)

In part 2, illustration of R to perform a two way factorial analysis of variance

Tip.
Before doing this illustration, do the illustration of using R to do a one-way analysis of variance. You can find this in the homework for week 9 (Unit 6 - Analysis of Variance, part 2 of 2).

Dataset used
`hers_640anova.xlsx`

Packages used:
`readxl, summarytools, ggplot2, tidyverse, knitr, car, emmeans`

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

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

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.

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`.

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

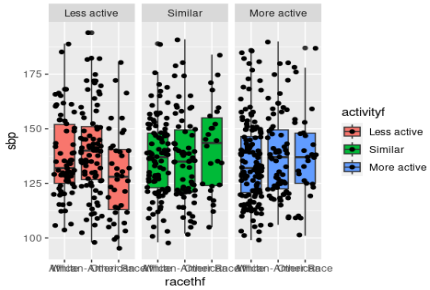
```
library(readxl)
source <- read_excel("hers_640anova.xlsx")
source <- as.data.frame(source)
str(source)

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

Prepare data for analysis of variance: Create factors and (recommended). Set reference levels explicitly.

```
library(tidyverse)
library(summarytools)

#  Factor I:  create factor variable racethf from 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")

ctable(x=source$physact, y=source$activityf, prop="n")                        # Check.

## 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
## --------- ----------- ------------- --------- ------------- -------
```

Always look at your data - 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
```

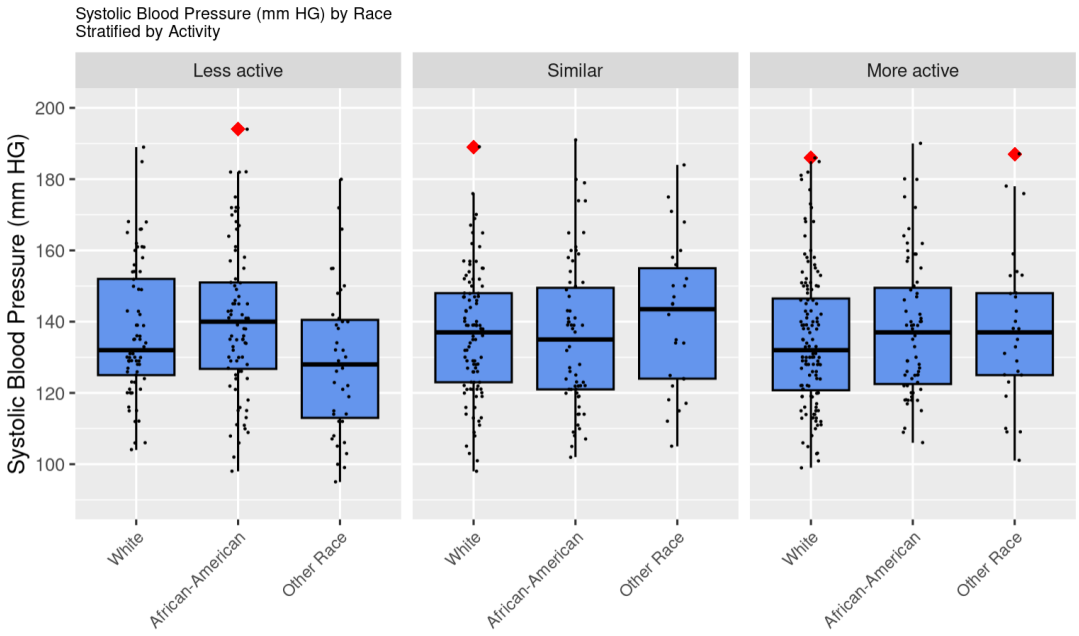


Basic graph does not look good. Needs fixing!

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

Always look at your data - With aesthetics for improved readability.
Y=sbp by X=racethf with stratification on Z = 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")
```



Obtain numerical descriptives: Custom!
Introduction to group_by( ) and summarise( )in the package {tidyverse}
And using kable( ) in the package {knitr} for pretty output.

```
library(tidyverse)
library(knitr)

mydescriptives2 <- source %>%
                  group_by(racethf,activityf) %>%
                  summarise(
                        n=n(),                                    # User specifies stats as separate options
                        mean=mean(sbp,na.rm=TRUE),                # Remove missing values using option 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

Fit model as analysis of variance. Show.
Introduction to function Anova( ) in package {car}.

```
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)


anova(m2_anova)

## 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

Anova(m2_anova, type="II")                                  # option type="II" in Anova() in {car} to obtain 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



Anova(m2_anova, type="III")                                 # option type="III" in Anova() in {car} to obtain 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
```

Fit model as regression using lm( ) and indicator variables. Show.
Introduction to mutate( ) and ifelse( ) in {tidyverse} to create 0/1 indicators

```
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)

anova(m2_regression)

## 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


summary(m2_regression)

## 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
```

TIP - In an analysis for 2 way factorial, the order of testing matters.
Test #1: Assess effect modification/interaction

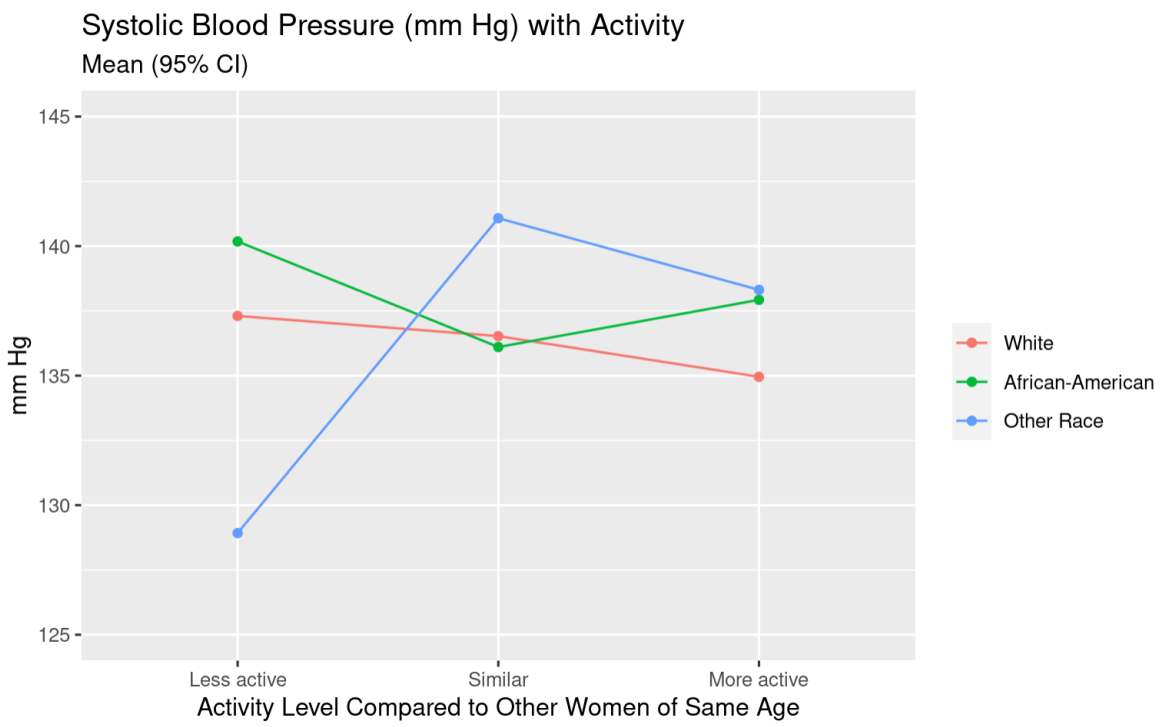
```
library(tidyverse)

# Test #1: Assess effect modification/ 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)

anova(full1, reduced1)

## 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



# Test #2:  Asess main effects
# Partial F-Test of Null:  No main effect of race in model with ZERO effect modification/interaction
full2 <- aov(sbp ~ activityf + racethf, data=source)
reduced2 <- aov(sbp ~ activityf, data=source)

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

# Test #2:  Asess main effects
# Partial F-Test of Null:  No main effect of activity in model with ZERO effect modification interaction
full3 <- aov(sbp ~ racethf + activityf, data=source)
reduced3 <- aov(sbp ~ racethf, data=source)

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
```

Report: Estimated means from previously saved anova model object
Current model is m2_anova: sbp ~ racethf + activityf + racethf:activityf
Introduction to emmeans( ) in package {emmeans}

```
library(emmeans)

# Predicted means of Y by factor=activity, stratified by racethf
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


# Predicted means of Y by factor=racethf, stratified by activityf
emm2 = emmeans::emmeans(m2_anova, specs = "racethf", by="activityf")
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
```

Report: Visualization of predictd means and 95% CI
NOTE - It is possible to do this using {emmeans}. I prefer using {ggplot2}
Introduction to using group_by( ) and summarise( ) in {tidyverse} to create a dataset for plotting

```
library(tidyverse)
library(ggplot2)

# get descriptives for plotting. Save as dataframe.
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.



# Plot of 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
```


Systolic Blood Pressure (mm Hg) with Activity
Mean (95% CI)
145
140
135
130
125
mm Hg
Less active
Similar
More active
Activity Level Compared to Other Women of Same Age
White
African-American
Other Race