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

**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<br>2 = african american<br>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*

## Preliminaries

**import excel data: lbw_anova**

```
library(readxl)
lbw_anova <- read_excel("lbw.xlsx")
lbw_anova <- as.data.frame(lbw_anova)
str(lbw_anova)

## 'data.frame':    189 obs. of  4 variables:
##  $ id  : num  85 86 87 88 89 91 92 93 94 95 ...
##  $ race: num  2 3 1 1 1 3 1 3 1 1 ...
##  $ bwt : num  2523 2551 2557 2594 2600 ...
##  $ ftv : num  0 3 1 2 0 0 1 1 1 0 ...
```

### #1.

State the analysis of variance model using notation $\mu$, $\alpha_i$, $\beta_j$, $(\alpha\beta)_{ij}$ and $\sigma^2$ as appropriate. Define all terms and constraints on the parameters.

**Answer:**

$X_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$, *where* $\varepsilon_{ijk} \sim Normal(0, \sigma_{error})$

*with* $i = 1, 2, 3$ *indexing race group*

$j = 1, 2$ *indexing visits to MD in* 1*st trimester (zero or at least one)*

$\mu$ = *population mean birthweight (g), over all groups*

$\alpha_i$ = *deviation from mean effect of race* $= i$, *with* $\sum_{i=1}^{3} \alpha_i = 0$

$\beta_j$ = *deviation from mean effect of visits to MD* $= j$ *with* $\sum_{j=1}^{2} = 0$

$(\alpha\beta)_{ij}$ = *extra deviation from mean (beyond main effects) with*

$$\sum_{i=1}^{3} (\alpha\beta)_{ij} = 0 \quad \text{and} \quad \sum_{i=j}^{2} (\alpha\beta)_{ij} = 0$$

### #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`**

```
library(tidyverse)

ready <- lbw %>%
         mutate(racef = recode_factor(race,                          # create factor var racef
                           "1" = "White",
                           "2" = "African American",
                           "3" = "Other")) %>%

         mutate(no_trimester1 = ifelse(ftv==0,1,0)) %>%              # numeric version (0/1)
         mutate(no_trimester1f= recode_factor(no_trimester1,         # factor version
                           "0" = "Visits",
                           "1" = "No visits"))

glimpse(ready)
```

```
## Rows: 189
## Columns: 7
## $ race           <dbl> 2, 3, 1, 1, 1, 3, 1, 3, 1, 1, 3, 3, 3, 3, 1, 1, 2, 1, 3…
## $ bwt            <dbl> 2523, 2551, 2557, 2594, 2600, 2622, 2637, 2637, 2663, 2…
## $ ftv            <dbl> 0, 3, 1, 2, 0, 0, 1, 1, 1, 0, 0, 1, 0, 2, 0, 0, 0, 3, 0…
## $ id             <dbl> 85, 86, 87, 88, 89, 91, 92, 93, 94, 95, 96, 97, 98, 99,…
## $ racef          <fct> African American, Other, White, White, White, Other, Wh…
## $ no_trimester1  <dbl> 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1…
## $ no_trimester1f <fct> No visits, Visits, Visits, Visits, No visits, No visits…
```

## #3.

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

```
library(tidyverse)

ready %>%                                                  # Summary by group using dplyr
  group_by(racef,no_trimester1f) %>%
  summarize(n=sum(!is.na(bwt)),                            # TIP - Use sum(!is.na()) to get number of complete observations
            mean=mean(bwt),
            sd=sd(bwt),
            min = min(bwt),
            P25 = quantile(bwt, 0.25),
            median = median(bwt),
            P75 = quantile(bwt, 0.75),
            max = max(bwt))
```

```
## # A tibble: 6 × 10
## # Groups:   racef [3]
##   racef          no_trimester1f     n  mean    sd   min   P25 median   P75   max
##   <fct>          <fct>          <int> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1 White          Visits            53 3176.  756.  1021 2663   3080  3770   4990
## 2 White          No visits         43 3013.  690.  1818 2417   3076  3622.  4238
## 3 African Ameri… Visits            12 2719.  620.  1701 2307.  2920  3047   3860
## 4 African Ameri… No visits         14 2720.  677.  1135 2381   2650. 3275.  3790
## 5 Other          Visits            24 2877.  651.  1588 2448   2792. 3308.  3997
## 6 Other          No visits         43 2763.  762.   709 2261   2863  3228.  4054
```

## #4.

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

```
fit_anova <- aov(bwt ~ racef + no_trimester1f + racef*no_trimester1f, data=ready)
anova(fit_anova)

## Analysis of Variance Table
##
## Response: bwt
##                       Df   Sum Sq Mean Sq F value   Pr(>F)
## racef                  2  5048361 2524181  4.9123 0.008354 **
## no_trimester1f         1   692771  692771  1.3482 0.247104
## racef:no_trimester1f   2   140324   70162  0.1365 0.872458
## Residuals            183 94033843  513846
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

KEY: With respect to the variability in birthweight (Y=bwt),
(1) There is NO statistically significant evidence of an interaction race x visits to MD (p-value = .87)
(2) There is NO statistically significant evidence of variations by whether or not the mother visited her Ob/GYN in the first trimester (p-value = .25)
(3) However, we do see statistically significant evidence of race group differences in mean birthweight (p-value = .008)

## #5.

This time, perform the two way analysis of variance as a regression.
Preliminaries: Create 0/1 indicator vars and interaction vars.

```
library(tidyverse)

ready <- ready %>%

         mutate(African_American = ifelse(race==2,1,0)) %>%          # 0/1 indicator African American

         mutate(Race_Other = ifelse(race==3,1,0)) %>%                # 0/1 indicator African American


         mutate(AfrAmer_novisits1 = African_American*no_trimester1) %>%    # interaction
         mutate(Other_novisits1 = Race_Other*no_trimester1)                # interaction

str(ready)

## 'data.frame':    189 obs. of  11 variables:
##  $ race             : num  2 3 1 1 1 3 1 3 1 1 ...
##  $ bwt              : num  2523 2551 2557 2594 2600 ...
##  $ ftv              : num  0 3 1 2 0 0 1 1 1 0 ...
##  $ id               : num  85 86 87 88 89 91 92 93 94 95 ...
##  $ racef            : Factor w/ 3 levels "White","African American",..: 2 3 1 1 1 3 1 3 1 1 ...
##  $ no_trimester1    : num  1 0 0 0 1 1 0 0 0 1 ...
##  $ no_trimester1f   : Factor w/ 2 levels "Visits","No visits": 2 1 1 1 2 2 1 1 1 2 ...
##  $ African_American : num  1 0 0 0 0 0 0 0 0 0 ...
##  $ Race_Other       : num  0 1 0 0 0 1 0 1 0 0 ...
##  $ AfrAmer_novisits1: num  1 0 0 0 0 0 0 0 0 0 ...
##  $ Other_novisits1  : num  0 0 0 0 0 1 0 0 0 0 ...
```

## #5.

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

```
fit_lm <- lm(bwt ~ African_American + Race_Other + no_trimester1 + AfrAmer_novisits1 + Other_novisits1,
             data=ready)
anova(fit_lm)

## Analysis of Variance Table
##
## Response: bwt
##                    Df   Sum Sq Mean Sq F value   Pr(>F)
## African_American    1  1520693 1520693  2.9594 0.087068 .
## Race_Other          1  3527668 3527668  6.8652 0.009527 **
## no_trimester1       1   692771  692771  1.3482 0.247104
## AfrAmer_novisits1   1   116469  116469  0.2267 0.634579
## Other_novisits1     1    23855   23855  0.0464 0.829646
## Residuals         183 94033843  513846
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(fit_lm)

##
## Call:
## lm(formula = bwt ~ African_American + Race_Other + no_trimester1 +
##     AfrAmer_novisits1 + Other_novisits1, data = ready)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2155.32  -513.32   -13.49   551.68  1813.68
##
## Coefficients:
##                   Estimate Std. Error t value            Pr(>|t|)
## (Intercept)        3176.32      98.46  32.259 <0.0000000000000002 ***
## African_American   -457.24     229.16  -1.995              0.0475 *
## Race_Other         -299.70     176.37  -1.699              0.0910 .
## no_trimester1      -163.67     147.12  -1.112              0.2674
## AfrAmer_novisits1   164.80     318.07   0.518              0.6050
## Other_novisits1      50.53     234.53   0.215              0.8296
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 716.8 on 183 degrees of freedom
## Multiple R-squared:  0.05886,    Adjusted R-squared:  0.03315
## F-statistic: 2.289 on 5 and 183 DF,  p-value: 0.04769
```

With respect to the variability in birthweight (Y=bwt), we learn a bit more with the regression approach:
(1) There is NO statistically significant evidence of an interaction race x visits to MD (p-values = .61 and .83)
(2) Compared to mothers of White race, mothers with race=OTHER have a mean birthweight that is estimated to be 299.70 grams lower (beta = -299.70) and the is marginally statistically significant (p-value = .09).
(3) Compared to mothers of White race, African American mothers have a mean birthweight that is estimated to be 457.24 grams lower (beta = -2457.24) and the is statistically significant (p-value = .0475).

## #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.

**Q6. partial F test of interactions - anova**

```
reduced1 <- aov(bwt ~ racef + no_trimester1f, data=ready)
full1 <- aov(bwt ~ racef + no_trimester1f + racef*no_trimester1f, data=ready)
anova(reduced1, full1)

## Analysis of Variance Table
##
## Model 1: bwt ~ racef + no_trimester1f
## Model 2: bwt ~ racef + no_trimester1f + racef * no_trimester1f
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1    185 94174167
## 2    183 94033843  2    140324 0.1365 0.8725
```

Controlling for the main effects of race and any visits to MD in the first trimester, there is no additional predictive significance of their interaction (Partial F test p-value = .87).

**Q6. partial F test of interactions - lm**

```
reduced2 <- lm(bwt ~ African_American + Race_Other + no_trimester1, data=ready)
full2 <- lm(bwt ~ African_American + Race_Other + no_trimester1 + AfrAmer_novisits1 + Other_novisits1,
            data=ready)
anova(reduced2,full2)
## Analysis of Variance Table
##
## Model 1: bwt ~ African_American + Race_Other + no_trimester1
## Model 2: bwt ~ African_American + Race_Other + no_trimester1 + AfrAmer_novisits1 +
##     Other_novisits1
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1    185 94174167
## 2    183 94033843  2    140324 0.1365 0.8725
```

Controlling for the main effects of race and any visits to MD in the first trimester, there is no additional predictive significance of their interaction (Partial F test p-value = .87).

## #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.

**Q7. predicted means - aov()**

```
library(emmeans)
# means over no_trimester1f, separately by racef
predicted_means7a = emmeans::emmeans(fit_anova, specs = "no_trimester1f", by="racef")
predicted_means7a

## racef = White:
##  no_trimester1f emmean    SE  df lower.CL upper.CL
##  Visits           3176  98.5 183     2982     3371
##  No visits        3013 109.3 183     2797     3228
##
## racef = African American:
##  no_trimester1f emmean    SE  df lower.CL upper.CL
##  Visits           2719 206.9 183     2311     3127
##  No visits        2720 191.6 183     2342     3098
##
## racef = Other:
##  no_trimester1f emmean    SE  df lower.CL upper.CL
##  Visits           2877 146.3 183     2588     3165
##  No visits        2763 109.3 183     2548     2979
##
## Confidence level used: 0.95
```

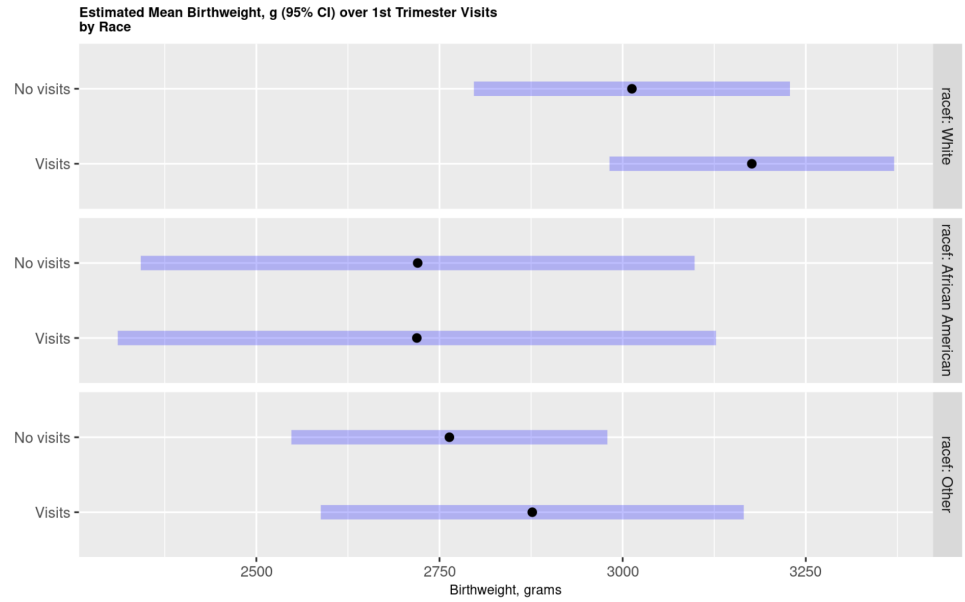
**Q7. predicted means - lm()**

```
library(tidyverse)

# Simplest is to now fit regression model with predictors as factor vars
fit_lm2 <- lm(bwt ~ factor(racef) + factor(no_trimester1f) + factor(racef)*factor(no_trimester1f),
              data=ready)

# Data frame of 6 groups:  Factor I (racef) at 3 levels x Factor II (no_trimester1f) at 2 levels
mygroups <- data.frame(
               racef = rep(factor(c("White", "African American","Other")),2),
               no_trimester1f = rep(factor(c("Visits", "No visits")),3)
             )

# Data frame of predicted means for the 6 groups
myhats <- as.data.frame(predict(fit_lm2, newdata = mygroups, interval = "confidence"))

# Combine group identification with predicted means
predicted_means7b <- cbind(mygroups,myhats)

# sort, rename variables for legibility and print
predicted_means7b <- predicted_means7b %>%
                     arrange(racef,no_trimester1f) %>%
                     rename(mean=fit,
                            'Lower 95% CI' = lwr,
                            'Upper 95% CI' = upr)
predicted_means7b
```

```
##              racef no_trimester1f     mean Lower 95% CI Upper 95% CI
## 2 African American      No visits 2720.214     2342.223     3098.206
## 5 African American         Visits 2719.083     2310.806     3127.361
## 6            Other      No visits 2763.488     2547.807     2979.169
## 3            Other         Visits 2876.625     2587.929     3165.321
## 4            White      No visits 3012.651     2796.970     3228.332
## 1            White         Visits 3176.321     2982.050     3370.592
```

**Supplement**
Plot of predicted means w 95% CI: From aov()

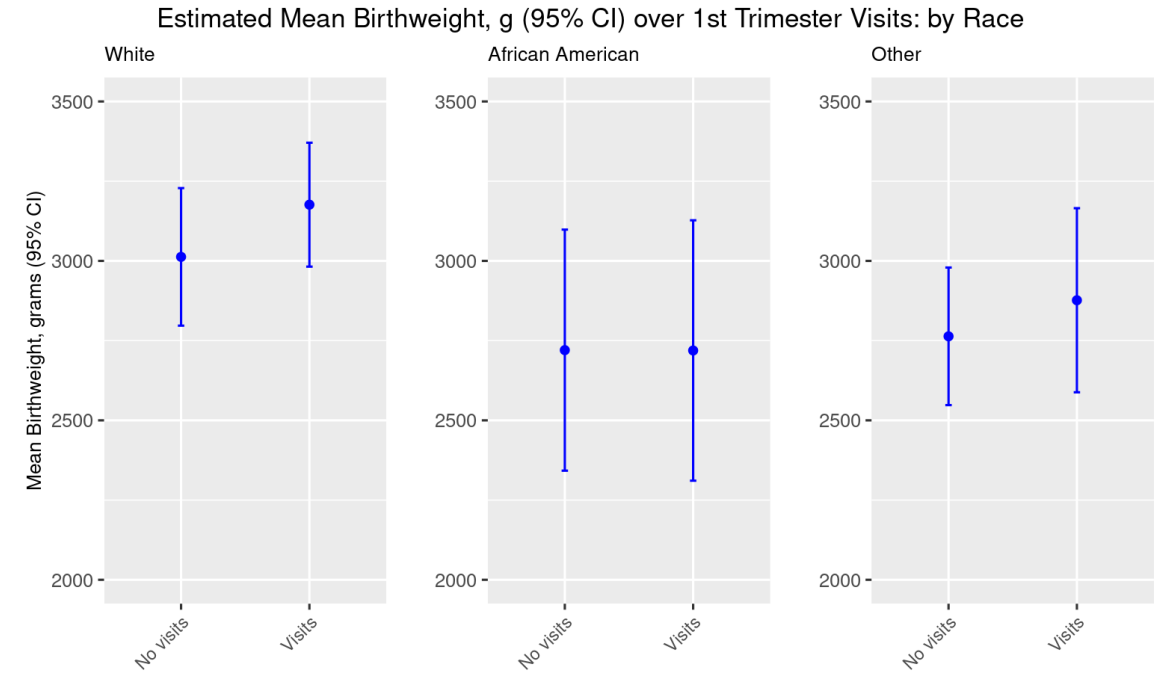
```
library(emmeans)
library(ggplot2)

# Get means over racef, separately by trimester1f
yhat_anova = emmeans::emmeans(fit_anova, specs = "no_trimester1f", by="racef")

# (extra).  Plot of means over no_trimester1f, stratified by racef
plot(yhat_anova) +
        labs(x="Birthweight, grams") +
        labs(y="") +
        ggtitle("Estimated Mean Birthweight, g (95% CI) over 1st Trimester Visits\nby Race") +
        theme(axis.title = element_text(size = 8),
                 plot.title = element_text(size = 8, face = "bold"))
```




**Supplement**
Plot of predicted means w 95% CI: From lm()

```
library(tidyverse)
library(ggplot2)
library(grid)
library(gridExtra)

# Simplest is to now fit regression model with predictors as factor vars
fit_lm2 <- lm(bwt ~ factor(racef) + factor(no_trimester1f) + factor(racef)*factor(no_trimester1f),
              data=ready)

# Data frame of 6 groups:  Factor I (racef) at 3 levels x Factor II (no_trimester1f) at 2 levels
mygroups <- data.frame(
               racef = rep(factor(c("White", "African American","Other")),2),
               no_trimester1f = rep(factor(c("Visits", "No visits")),3)
             )

# Data frame of predicted means for the 6 groups
myhats <- as.data.frame(predict(fit_lm2, newdata = mygroups, interval = "confidence"))


# Combine group identification with predicted means
mydata <- cbind(mygroups,myhats)


# (extra).  Plot of means over no_trimester1f, separately by racef
panel1 <- mydata %>% filter(racef=="White")
p1 <- ggplot(data=panel1) +
      aes(x=no_trimester1f, y=fit) +
      geom_errorbar(aes(ymin=lwr,
                        ymax=upr),
                        width=.05,
                        color="blue")  +
      geom_point(color="blue") +
      scale_y_continuous(limits=c(2000,3500), breaks=seq(2000,3500,500)) +    # Be sure to use SAME y-axis scale for all
      ggtitle("White") +
      xlab(" ")+
      ylab("Mean Birthweight, grams (95% CI)") +                               # Y-axis label for left most panel only
      theme(axis.text.x = element_text(size = 8, angle=45, vjust=1, hjust=1),
               axis.title = element_text(size = 9),
               plot.title = element_text(size = 9))
```

```
panel2 <- mydata %>% filter(racef=="African American")
p2 <- ggplot(data=panel2) +
      aes(x=no_trimester1f, y=fit) +
      geom_errorbar(aes(ymin=lwr,
                        ymax=upr),
                        width=.05,
                        color="blue")  +
      geom_point(color="blue") +
      scale_y_continuous(limits=c(2000,3500), breaks=seq(2000,3500,500)) +    # Be sure to use SAME y-axis scale for all
      ggtitle("African American") +
      xlab(" ")+
      ylab(" ") +                                                              # Y-axis label here would be too busy
      theme(axis.text.x = element_text(size = 8, angle=45, vjust=1, hjust=1),
               axis.title = element_text(size = 9),
               plot.title = element_text(size = 9))


panel3 <- mydata %>% filter(racef=="Other")
p3 <- ggplot(data=panel3) +
      aes(x=no_trimester1f, y=fit) +
      geom_errorbar(aes(ymin=lwr,
                        ymax=upr),
                        width=.05,
                        color="blue")  +
      geom_point(color="blue") +
      scale_y_continuous(limits=c(2000,3500), breaks=seq(2000,3500,500)) +    # Be sure to use SAME y-axis scale for all
      ggtitle("Other") +
      xlab(" ")+
      ylab(" ") +                                                              # Y-axis label here would be too busy
      theme(axis.text.x = element_text(size = 8, angle=45, vjust=1, hjust=1),
               axis.title = element_text(size = 9),
               plot.title = element_text(size = 9))

gridExtra::grid.arrange(p1, p2, p3, ncol=3,
                        top=textGrob("Estimated Mean Birthweight, g (95% CI) over 1st Trimester Visits: by Race"))
```



## 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<br>2 = African American<br>3 = Other |
| physact | Comparative ("compared to other women your age") physical activity | Numeric | 1 = much less active<br>2 = somewhat less active<br>3 = about as active<br>4 = somewhat more active<br>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**

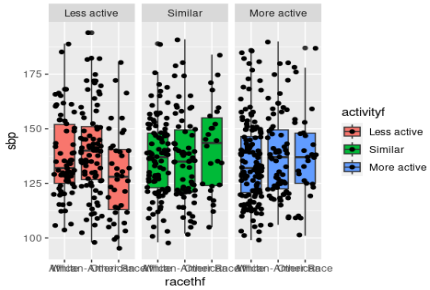
```
library(readxl)
source <- read_excel("hers_640anova.xlsx")
source <- as.data.frame(source)
str(source)

## '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
```

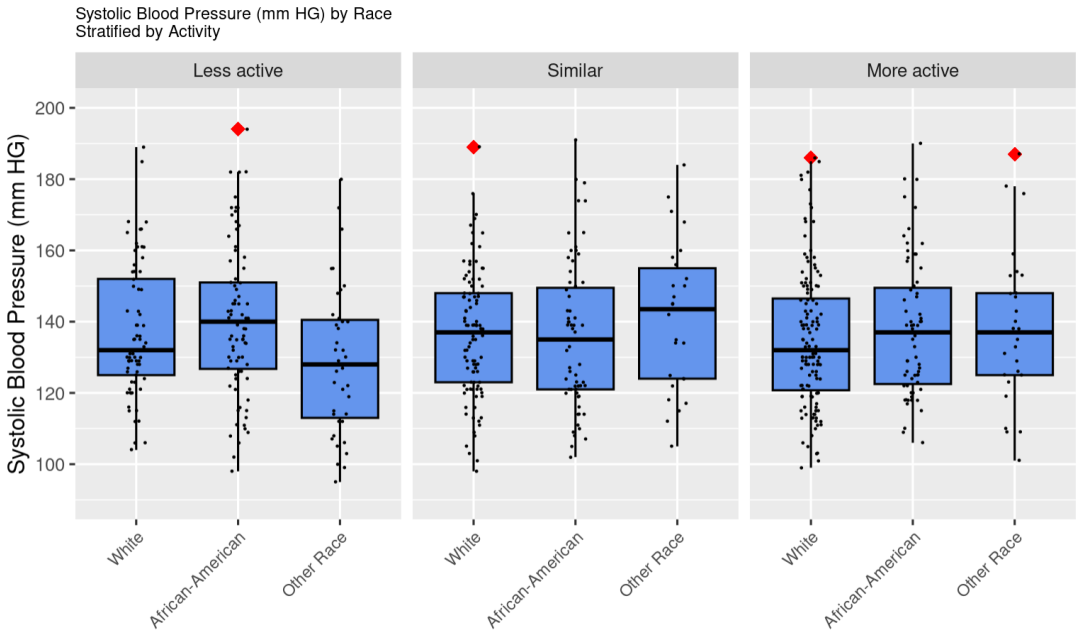



Basic graph does not look good. Needs fixing!

```
  #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**

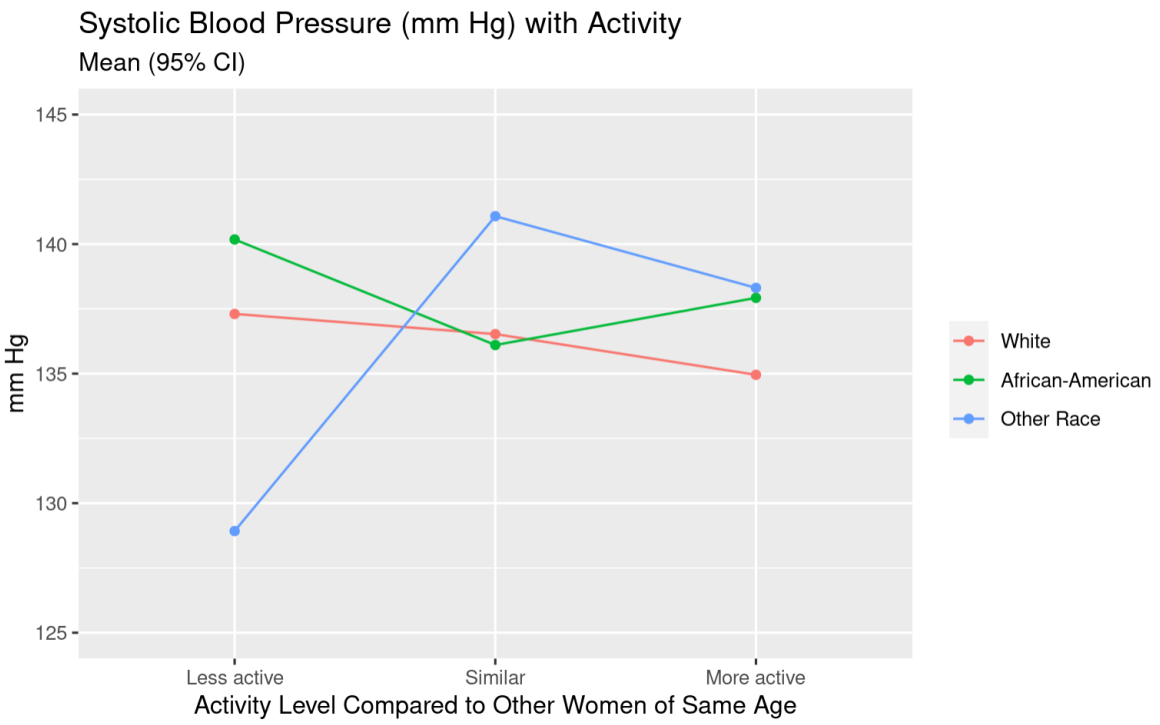
```
library(tidyverse)

# 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