

## Unit 6 – Analysis of Variance Practice Problems (1 of 2)

### Solutions

**Before you begin.** Download from the course website  
`anova_infants.xlsx`

Zelazo et al. (1972) studied the variability in age at first walking in infants. 24 infants were randomly assigned to four groups of equal sample size (6 infants per group), with groups defined by method of reinforcement of walking: (1) active (2) passive (3) no exercise; and (4) 8 week control. The outcome variable measured was age at first walking, in months. The following table lists the study data, by group.

**Table – Study Data of Zelazo et al (1972), n=24:**

Active Group	Passive Group	No-Exercise Group	8 Week Control
9.00	11.00	11.50	13.25
9.50	10.00	12.00	11.50
9.75	10.00	9.00	12.00
10.00	11.75	11.50	13.50
13.00	10.50	13.25	11.50
9.50	15.00	13.00	12.35

Source: Zelazo et al (1972) “Walking” in the newborn. *Science* 176: 314-315.

### Data dictionary/Codebook:

Variable	Label	Type	Coding
<b>group</b>	Group	numeric	1 = active 2 = passive 3 = noex 4 = control
<b>age</b>	Age, months	numeric	Continuous, months
<b>I_active</b>	Indicator group = “active”	numeric	1 if group = 1 (“active”) 0 otherwise
<b>I_passive</b>	Indicator group=”passive”	numeric	1 if group = 2 (“passive”) 0 otherwise
<b>I_noex</b>	Indicator group = “noex”	numeric	1 if group = 3 (“noex”) 0 otherwise

#1.

**Deviation from means.** State the analysis of variance model using deviation from means notation  $\mu$  and  $\tau_i$  and  $\sigma^2$  as appropriate. Define all terms and constraints on the parameters.

Answer:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \text{ where } \varepsilon_{ij} \sim N(0, \sigma^2) \text{ and } \sum_{i=1}^4 \tau_i = 0$$

$i = 1, 2, \dots, K$  indexes method of reinforcement group;

$K = \text{number of groups} = 4$

$j=1, 2, \dots, n_i=6$  indexes infant within group;

$\mu = \text{population mean age at first walking, over all groups}$

$\mu_i = \text{mean age at first walking for infants in group "i"}$

$\tau_i = [\mu_i - \mu]$

$Y_{ij} = \text{observed age at first walking for the } j\text{th infant in group "i"}$

$H_0: \tau_1=0, \tau_2=0, \tau_3=0, \text{ and } \tau_4=0$

$H_A: \text{At least one } \tau_i \neq 0$

Import excel data

```
library(readxl)
infants <- read_excel("anova_infants.xlsx")
infants <- as.data.frame(infants)
str(infants)
```

```
## 'data.frame':    24 obs. of  7 variables:
## $ group      : num  1 1 1 1 1 1 2 2 2 2 ...
## $ age        : num  9 9.5 9.75 10 13 ...
## $ passive    : num  0 0 0 0 0 0 1 1 1 1 ...
## $ noex       : num  0 0 0 0 0 0 0 0 0 0 ...
## $ I_active   : num  1 1 1 1 1 1 0 0 0 0 ...
## $ I_passive  : num  0 0 0 0 0 0 1 1 1 1 ...
## $ I_noex     : num  0 0 0 0 0 0 0 0 0 0 ...
```

Create factor variable groupf

```
library(tidyverse)

infants <- infants %>%
  mutate(groupf= recode_factor(group,
                                "1" = "active",
                                "2" = "passive",
                                "3" = "noex",
                                "4" = "control"))
```

```
table(infants$group,infants$groupf)
##      active passive noex control
## 1         6         0         0         0
## 2         0         6         0         0
## 3         0         0         6         0
## 4         0         0         0         6
```

# check group x groupf crosstab

Preliminary: descriptives by group

```
library(summarytools) # stby() in package {summarytools}

with(infants,
  stby(data = age,
    INDICES = groupf,
    FUN = descr,
    stats = c("n.valid", "mean", "sd", "min", "med", "max"),
    transpose=TRUE)) # with(DATAFRAMEName,
# stby(data=OUTCOMEVariable,
# INDICES=GROUPVariable, must be factor
# user chooses statistics to show.

## Descriptive Statistics
## age by groupf
## Data Frame: infants
## N: 6
##
##      N.Valid   Mean   Std.Dev   Min   Median   Max
## -----
## active      6.00   10.12     1.45    9.00     9.62   13.00
## passive     6.00   11.38     1.90   10.00    10.75   15.00
## noex        6.00   11.71     1.52    9.00    11.75   13.25
## control     6.00   12.35     0.86   11.50    12.18   13.50
```

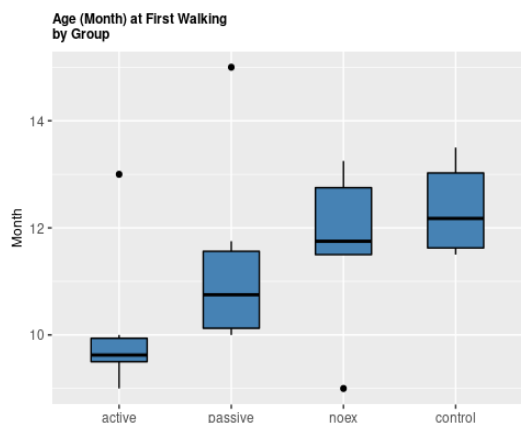
## #2.

By any means you like, produce a side by side box plot showing the distribution of age at first walking, separately for each of the 4 groups.

Q2: side-by-side plot of outcome (age) by predictor (groupf)

```
library(ggplot2)

ggplot(data=infants) +
  aes(x=groupf) +
  aes(y=age) +
  geom_boxplot(color="black",
    fill="steelblue",
    width=0.5) +
  #geom_jitter(shape = 18,
    #color = "black",
    #position = position_jitter(0.03)) +
  ggtitle("Age (Month) at First Walking\nby Group") + # Optional aesthetics: titles and axis labels
  xlab(" ") +
  ylab("Month") +
  theme(axis.text = element_text(size=9),
    axis.title = element_text(size=9),
    plot.title = element_text(size=9, face="bold")) # Optional aesthetics: font size selections
```



Interpretation

- In these data, first walking occurs earlier when infants are reinforced
- Distributions differ markedly with respect to variability with greatest seen among infants in the passive group and smallest among infants in the control group

#3.

By any means you like, obtain the entries of the analysis of variance table for this one way analysis of variance. Use your computer output (or excel work or hand calculations or whatever) to complete the following table:

Source	df	Sum of Squares SSQ	Mean Square MSQ	F-Statistic	p-value
Between Groups	3	15.74	5.25	2.40	.10
Within Groups	20	43.69	2.18		
Total, corrected	23	59.43			

Q3: One way analysis of variance - table

```
fit_anova <- aov(age ~ groupf, data=infants)
anova(fit_anova)
```

```
## Analysis of Variance Table
```

```
##
```

```
## Response: age
```

```
##          Df Sum Sq Mean Sq F value Pr(>F)
```

```
## groupf    3  15.74  5.2468  2.4018 0.09787 .
```

```
## Residuals 20  43.69  2.1845
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#4.

Write a 2-5 sentence report of your description and hypothesis test findings using language as appropriate for a client who is intelligent but is not knowledgeable about statistics. Consider including a figure and/or table that you think is appropriate.

In this sample, the data suggest a trend towards earlier age at first walking with increasing reinforcement and placement. The median age at first walking is greatest among controls (12.35 months) and lowest among infants in the “active” group (10.13 months); see also the box plots. Tests of statistical significance were limited to the overall F test for group differences and this did not achieve statistical significance (p-value = .10), possibly due to the small sample sizes (6 in each group).

Interestingly, examination of the data also suggests that the variability in age at first walking differed, depending on the intervention received. The variability was greater in the three intervention groups (“active”, “passive”, “no exercise”) compared to in the “control” group; this was not statistically significant however (p-value = .45).

Further study, utilizing larger sample sizes and additional hypothesis tests to investigate trend are needed.

#5.

**Reference cell coding** Repeat your analysis, this time using what you learned in Unit 5 - Normal Theory Regression. Specifically, using appropriately defined indicator variables, perform a multivariable linear regression analysis of these same data! Use your computer output to complete the following table:

Source	df	Sum of Squares SSQ	Mean Square MSQ	F-Statistic	p-value
Due Model	3	15.74	5.25	2.40	.10
Due Error (residual)	20	43.69	2.18		
Total, corrected	23	59.43			

Q5: Multiple predictor regression - Modeling categorical using user created design variables

```
fit_lm1 <- lm(age ~ I_active + I_passive + I_noex, data=infants) # Categorical levels modeled explicitly using 0/1
anova(fit_lm1)
## Analysis of Variance Table
##
## Response: age
##           Df Sum Sq Mean Sq F value Pr(>F)
## I_active   1 12.793  12.7934   5.8565 0.02516 *
## I_passive   1   1.712   1.7117   0.7836 0.38656
## I_noex      1   1.235   1.2352   0.5654 0.46083
## Residuals 20 43.690   2.1845
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q5: Multiple predictor regression - Modeling categorical as factor variable

```
infants$groupf <- relevel(infants$groupf, ref = "control") # relevel() with option ref= to set reference

fit_lm2 <- lm(age ~ factor(groupf), data=infants) # factor( ) to model categorical levels under the hood
summary(fit_lm2)
##
## Call:
## lm(formula = age ~ factor(groupf), data = infants)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.7083 -0.8500 -0.2792  0.5062  3.6250
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    12.3500     0.6034  20.468 0.0000000000000694 ***
## factor(groupf)active    -2.2250     0.8533  -2.607    0.0169 *
## factor(groupf)passive   -0.9750     0.8533  -1.143    0.2667
## factor(groupf)noex     -0.6417     0.8533  -0.752    0.4608
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.478 on 20 degrees of freedom
## Multiple R-squared:  0.2649, Adjusted R-squared:  0.1546
## F-statistic: 2.402 on 3 and 20 DF, p-value: 0.09787

anova(fit_lm2)
## Analysis of Variance Table
##
## Response: age
##           Df Sum Sq Mean Sq F value Pr(>F)
## factor(groupf)  3  15.74   5.2468   2.4018 0.09787 .
## Residuals      20 43.69   2.1845
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#6.

**Deviation from means** and **reference cell coding** are equivalent! Using your output from your two analyses (1<sup>st</sup>-analysis of variance, 2<sup>nd</sup> – regression), obtain the predicted mean of Y =age at first walking twice in two ways.

	Prediction Using One Way Analysis of Variance	Prediction Using Multiple Linear Regression
Active	10.125	$\hat{\mu}_1 = (\hat{\beta}_0 + \hat{\beta}_1) = 12.35 - 2.225 = 10.125$
Passive	11.375	$\hat{\mu}_2 = (\hat{\beta}_0 + \hat{\beta}_2) = 12.35 - 0.97 = 11.38$
No-Exercise	11.71	$\hat{\mu}_3 = (\hat{\beta}_0 + \hat{\beta}_3) = 12.35 - 0.64 = 11.71$
Control (referent)	12.35	$\hat{\mu}_4 = \hat{\beta}_0 = 12.35$

Q6: Obtain predicted means for deviation from means coding ANOVA

```
aactive <- predict(fit_anova,data.frame(groupf="active"))
apassive <- predict(fit_anova,data.frame(groupf="passive"))
anoex <- predict(fit_anova,data.frame(groupf="noex"))
acontrol <- predict(fit_anova,data.frame(groupf="control"))
```

```
anames <- c("Active", "Passive", "NoEx", "Control")
ahat <- c(aactive,apassive,anoex,acontrol)
means.anova <- data.frame(anames,ahat)
means.anova
```

```
##      anames      ahat
## 1 Active 10.12500
## 2 Passive 11.37500
## 3 NoEx 11.70833
## 4 Control 12.35000
```

Q6: Obtain predicted means from reference cell coding REGRESSION

```
active <- predict(fit_lm1,data.frame(I_active=1,I_passive=0,I_noex=0))
passive <- predict(fit_lm1,data.frame(I_active=0,I_passive=1,I_noex=0))
noex <- predict(fit_lm1,data.frame(I_active=0,I_passive=0,I_noex=1))
control <- predict(fit_lm1,data.frame(I_active=0,I_passive=0,I_noex=0))
```

```
names <- c("Active", "Passive", "NoEx", "Control")
yhat <- c(active,passive,noex,control)
means.regression <- data.frame(names,yhat)
means.regression
```

```
##      names      yhat
## 1 Active 10.12500
## 2 Passive 11.37500
## 3 NoEx 11.70833
## 4 Control 12.35000
```

# **Supplement - Learn R (nothing to turn in)** Practice with R to perform a one way analysis of variance

## Dataset used

hers\_640anova.xlsx

## Packages used:

ggplot2, summarytools, tidyverse, HH, car, MASS, readxl

## 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

```
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 ...
```

Categorical predictors in R must be type factor. As you like, set the reference level explicitly

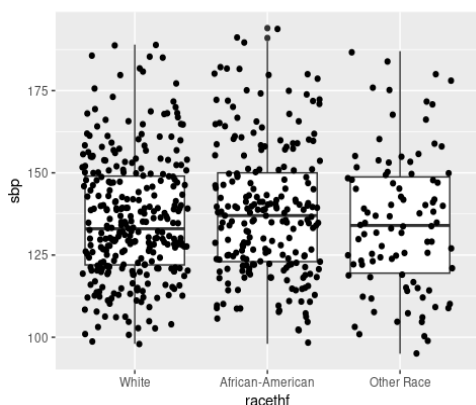
```
source$racethf <- factor(source$raceth, # Tip: Set explicitly
  levels=c(1,2,3),
  labels=c("White", "African-American", "Other Race"))

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

Always look at your data!

```
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



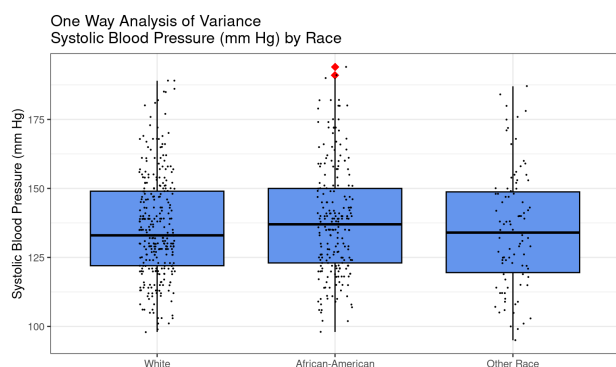
```
library(ggplot2)

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

  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) +

  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

#### Obtain numerical descriptives

```
library(summarytools)

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

with(source,
      stby(data = sbp,
           INDICES =racethf,
           FUN = descr,
           stats=c("n.valid", "pct.valid", "mean", "sd", "min", "max"),
           #stats=c("common"),
           transpose=TRUE))

# with(dataframe,
# # data = yvar
# # INDICES = factor var
# # user chooses statistics to show
# # NOT RUN: another set to show
# # display descriptives horizontally
```

#### ## 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)

Fit model as an analysis of variance. Show.

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

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
## Residuals 609 231671   380.41
##
summary(m1_anova) # summary(MODELFIT) to show results
##      Df Sum Sq Mean Sq F value Pr(>F)
## racethf    2    871   435.50   1.145  0.319
## Residuals 609 231671   380.4
##
anova( ) provides slightly more information than summary( )
```

Overall F test (Null: equality of means) does NOT reject null (p=.32)

Fit model as a regression. Show.

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

anova(m1_regression)
## Analysis of Variance Table
##
## Response: sbp
##      Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(racethf)    2    871   435.50   1.1448  0.319
## Residuals           609 231671   380.41
##
summary(m1_regression) # summary(MODELFIT) to show results
## Call:
## lm(formula = sbp ~ as.factor(racethf), 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(racethf)2    2.2206     1.7358    1.279    0.201
## as.factor(racethf)3   -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
```

Not surprising: no effect of racethf = 2  
Similarly, no effect of racethf = 3

Not surprising that R-squared is tiny!

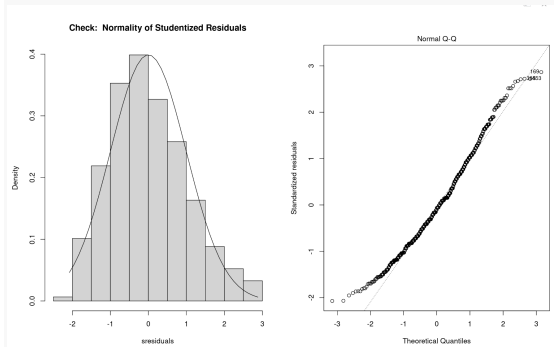
Regression diagnostics: normality of residuals - visualization

```
library(MASS)
```

```
sresiduals <- studres(m1_anova)
par(mfrow = c(1,2))
hist(sresiduals,
     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)

# Null: studentized residuals are Normal(0,1)
# set graph to be 2 panes (1 row, 2 col)
# histogram of sresiduals (plot density not freq)

# 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)
```

Regression diagnostics: normality of residuals - hypothesis test of Null: normality

```
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
```

Regression diagnostics: constant variance - hypothesis test of Null: homogeneity

```
library(HH)
library(car)

# hov() for Brown-Forsyth in package {HH}
# LeveneTest() in package {car}

# null: constant variance, all is well
bartlett.test(sbp ~ racethf, data=source)
## Bartlett test of homogeneity of variances
##
## data: sbp by racethf
## Bartlett's K-squared = 3.1766, df = 2, p-value = 0.07043

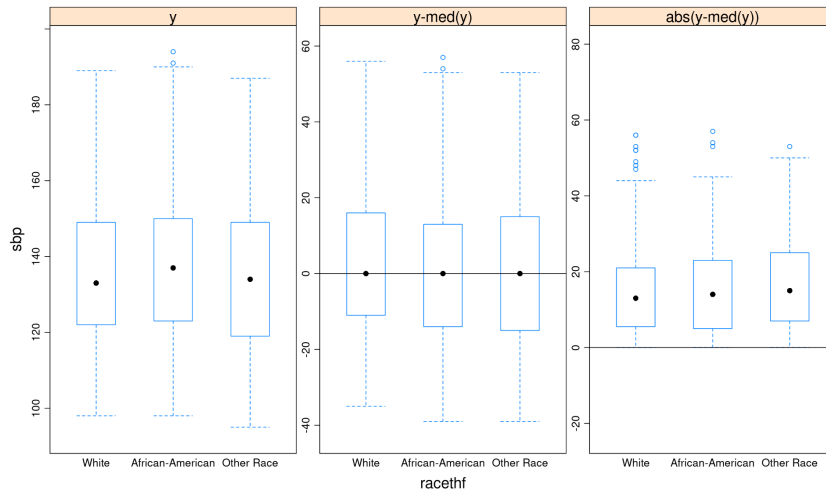
Test of Null: constant variance is NOT rejected (good!)
```

Regression diagnostics: constant variance - visualization

```
library(HH)
```

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

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



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