Two-Way ANOVA

By: Robert Russ

Regis University

Description of Data Set

The dataset given displays incomes (in $1000's USD) of males and females based on regions (north, south, east, west) that researcher wants to investigate. The two factors or predictor variables are region and gender while the response variable is income. The first column of the data has categorical data, the second column has the incomes of males and third column has incomes of females. We want to investigate the incomes of males and females in these different regions by using a two-way ANOVA analysis.

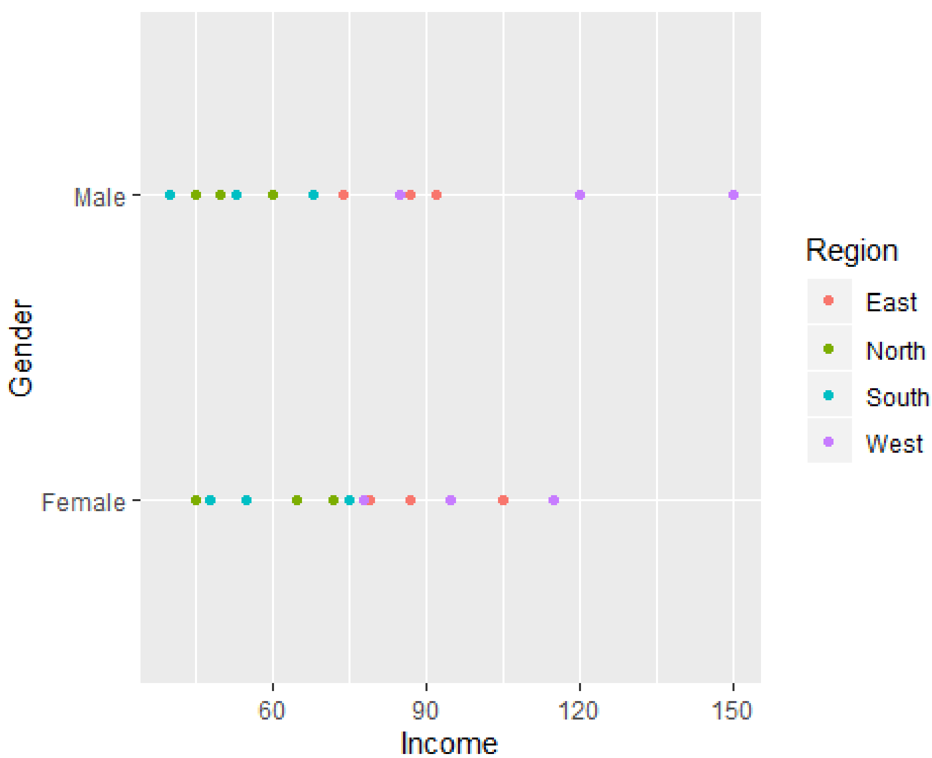
We will investigate the visualizations of the data to give us a better idea what the data is telling us. Prior to creating visualizations, I noticed the dataset was in a format not suitable for ANOVA analysis. I manipulated it by creating three columns. The first column is Region a categorical predictor variable, second column is Gender a categorical predictor variable, and third column is Income a continuous response variable. This enables me to fit a model to the dataset.

```
# loaded necessary libraries
library(car)

library(ggplot2)

library(ggpubr)

library(multcomp)

# load data into income object
income <- read.csv("income.csv")
head(income)

##   Region Gender Income
## 1  North Female     45
## 2  North Female     72
## 3  North Female     65
## 4  North   Male     50
## 5  North   Male     60
## 6  North   Male     45
```

```
str(income)

## 'data.frame':    24 obs. of  3 variables:
##  $ Region: Factor w/ 4 levels "East","North",..: 2 2 2 2 2 2 3 3 3 3 ...
##  $ Gender: Factor w/ 2 levels "Female","Male": 1 1 1 2 2 2 1 1 1 2 ...
##  $ Income: int  45 72 65 50 60 45 55 75 48 40 ...
```

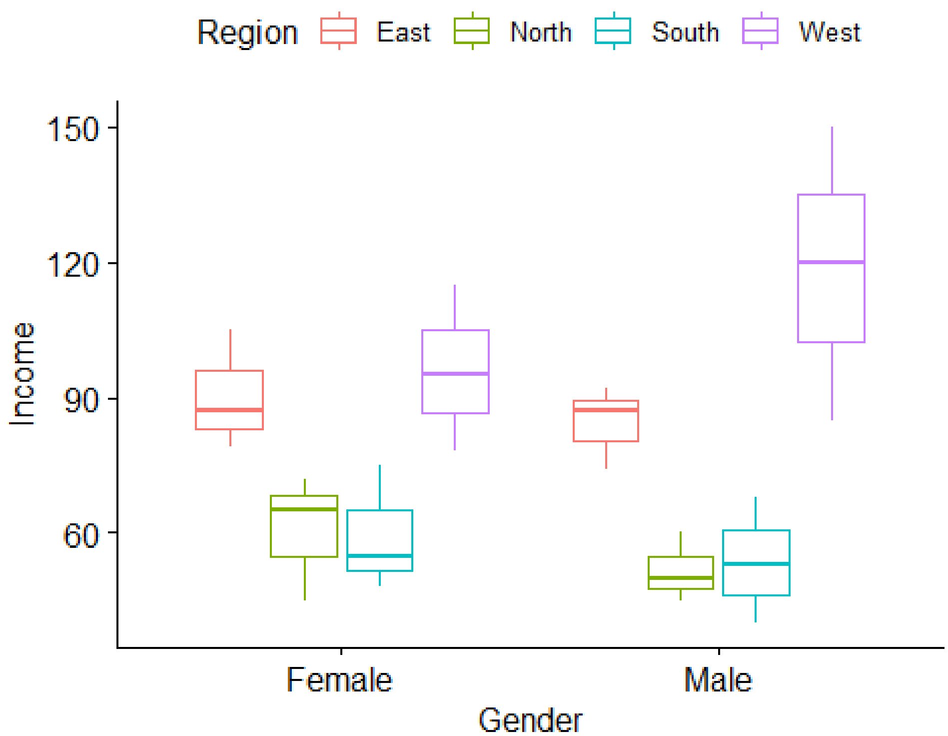
With the command of str(income), I noticed Region and Gender are recognized as factors confirms they are categorical, so R converted the four regions and gender into numbers. R recognized the Income variable as having integers which confirms it is continuous.

```
# visualize the data
ggplot(income, aes(Income, Gender, colour =
                    Region)) + geom_point()
```



In the dot plot above, we can see the income levels between the two genders in the North and South are lower than in the East and West. We can observe the income levels for female and males are close to being equal. Base on the plot, my claim is income levels depend more on Regions rather than Gender. I will validate or negate this claim later in my analysis.

```
ggboxplot(income, x = "Gender", y = "Income", color = "Region")
```



I plotted box plots to check the data for outliers and to check the data for balance. We can see in the East, North, and South Regions are balanced and not indication of outliers. The West Region shows an outlier and an imbalance or data.

Two-Way ANOVA Analysis

The two-way ANOVA analysis will start with the stating the null and alternative hypothesis for the main effects. The null hypothesis is the means of the different regions of given genders are not different from each other. The alternative hypothesis is that the different regions are different from each other. (Curley, 2017)

$$H_O : \mu_1 = \mu_2 = \ldots = \mu_K$$

H_A : Not H_O (Curley, 2017)

```
# execute the ANOVA model
income.mod1 <- aov(Income ~ Gender + Region, data = income)
summary(income.mod1)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## Gender       1      1       1   0.004    0.953    
## Region       3  11225    3742  12.706 8.69e-05 ***
## Residuals   19   5595     294                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to our results, we can conclude that the main effect of Gender is not statistically significant due to the p-value = 0.953 > 0.05; therefore, we will not reject the null hypothesis and say there is no difference in means of Gender. On the other hand, we see from our results that Region is statistically significant due to the p-value = 8.69e-05 < 0.05; therefore, we will reject the null hypothesis and say at least two of the means are different Regions. This also means Income depends on Region more than Gender.

```
income.mod2 <- aov(Income ~ Gender*Region, data = income)
summary(income.mod2)
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)    
## Gender         1      1       1   0.004 0.952873    
## Region         3  11225    3742  12.946 0.000151 ***
## Gender:Region  3    971     324   1.120 0.370514    
## Residuals     16   4625     289                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the above results, we can see the test for interaction effect (Gender:Region). In the next section, we will discuss the conclusions of the interaction effect result.

Interaction Effects

We will test the interaction effect with the following hypotheses to see whether Region and Gender have a linear relationship or correlated. The null hypothesis is there is no interaction effect between Region and Gender on Income. The alternative hypothesis is there is an interaction effect (Curley, 2017). Curley (2017) displayed the null and alternative as the illustration below:

H_0: No interaction

H_A: Interaction

The p-value for the interaction effect equals 0.370514 which is greater than 0.05. This means we will not reject the null hypothesis and say there is not interaction affect between Region and Gender. Since there is not interaction affect, we can say the main effect of Region is significant.

I wanted to do more investigation into the differences of means for Region, so I created several tables to see the interaction effect between the Regions.

```
model.tables(income.mod2, type = "means", se = TRUE)

## Tables of means
## Grand mean
##
## 76.79167
##
##  Gender
## Gender
## Female   Male
##  76.58  77.00
##
##  Region
## Region
##   East  North  South   West
##  87.33  56.17  56.50 107.17
##
##  Gender:Region
##         Region
## Gender   East   North  South  West
##   Female  90.33  60.67  59.33  96.00
##   Male    84.33  51.67  53.67 118.33
##
## Standard errors for differences of means
##         Gender Region Gender:Region
```

```
##          6.941  9.816        13.881
## replic.     12      6             3
```

We can see the means between Genders is significantly low almost none. This means person's income will not depend on his or her gender. We can see the means for the Regions are high. This means a person's income is dependent on the Region they work.

```
###Performing multiple pariwise-comparisons###
TukeyHSD(income.mod2, which = "Region")

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Income ~ Gender * Region, data = income)
##
## $Region
##                    diff       lwr       upr     p adj
## North-East  -31.1666667 -59.24947 -3.083863 0.0270886
## South-East  -30.8333333 -58.91614 -2.750530 0.0289698
## West-East    19.8333333  -8.24947 47.916137 0.2215604
## South-North   0.3333333 -27.74947 28.416137 0.9999853
## West-North   51.0000000  22.91720 79.082804 0.0004633
## West-South   50.6666667  22.58386 78.749470 0.0004953

summary(glht(income.mod2, linfct = mcp(Region = "Tukey")))

## Warning in mcp2matrix(model, linfct = linfct): covariate interactions foun
d
## -- default contrast might be inappropriate

##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = Income ~ Gender * Region, data = income)
##
## Linear Hypotheses:
##                    Estimate Std. Error t value Pr(>|t|)
## North - East == 0   -29.667     13.881  -2.137   0.1838
## South - East == 0   -31.000     13.881  -2.233   0.1567
## West - East == 0      5.667     13.881   0.408   0.9763
## South - North == 0   -1.333     13.881  -0.096   0.9997
## West - North == 0    35.333     13.881   2.545   0.0907 .
## West - South == 0    36.667     13.881   2.641   0.0758 .
## ---
```

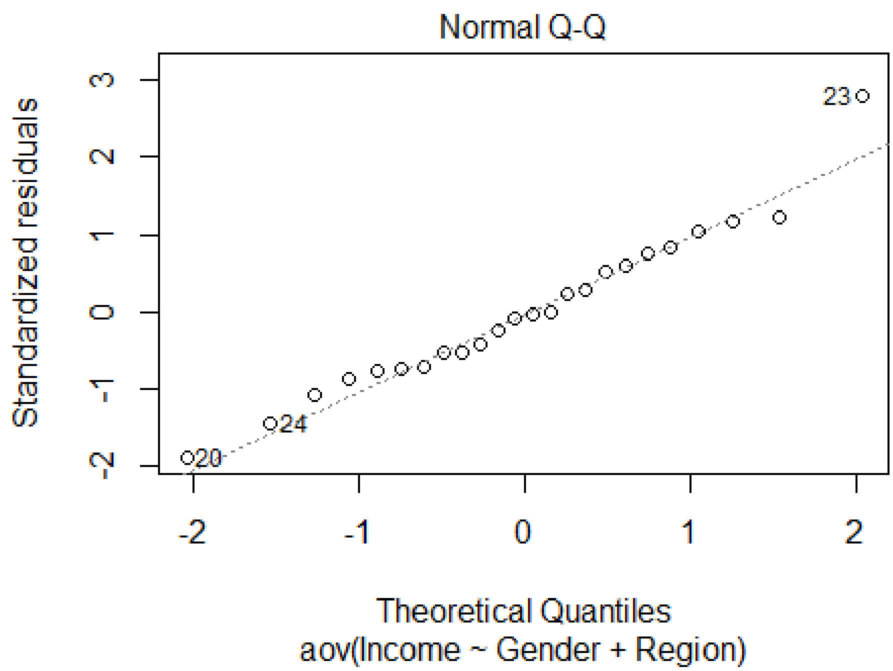
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

We will investigate the relationships among the Regions. I use the Tukey Honestly Significant Difference and General Linear Hypothesis tests. We can see from the Tukey HSD results North-East, South-East, West-North, and West-South have statistically significant results with p-values less than 0.05 which means these relationships have different means. For the GLH test, we see West-North and West-South are statistically significant with p-values less than 0.05. We can conclude that West-North and West-South means are different.

Conclusion

According to our analysis, we can say the main effects of Regions is significant, and there was not interaction effect between Region and Gender. In order to validate our model, we need to check two assumptions about two-way ANOVA.

```
#Assumption 1: The model residuals are normally distributed
plot(income.mod1, 2)
```

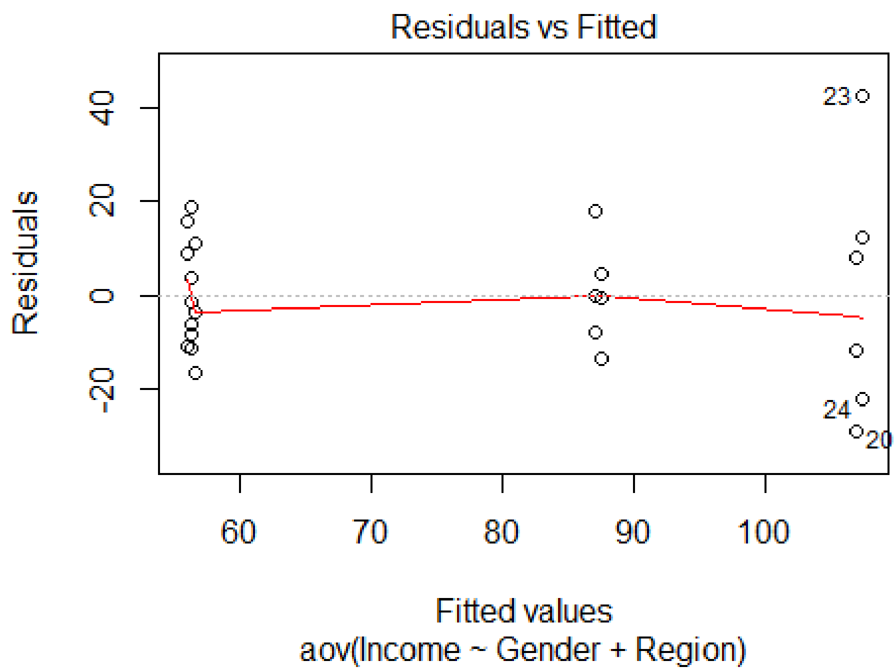



```
aov_residuals <- residuals(object = income.mod1)
shapiro.test(aov_residuals)

##
##  Shapiro-Wilk normality test
##
## data:  aov_residuals
## W = 0.96971, p-value = 0.6598
```

The first assumption I used the Normal Q-Q plot and Shapiro-Wilk normality test to verify the residuals are normally distributed. According to the graph, there aren't many points moving away from the line which means normality exists. According to the Shapiro-Wilk test, we see the p-value = 0.6598 greater than 0.05 which means the residuals are not significantly different from the normal distribution. We can assume the residuals are normally distributed.

Assumption 1 holds.

```
# Assumption 2: Homogenity of variance of the groups
plot(income.mod1, 1)
```




```
leveneTest(Income ~ Gender * Region, data = income)

## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  7   0.714 0.6617
##       16
```

The second assumption I used the Residuals versus Fitted plot and Levene's Test for Homogeneity of Variance test to verify the variances are equal. According to the graph, we can see the estimated regression line corresponds to the residual = 0 line. According to the Levene's test, we see the p-value = 0.6617 greater than 0.05 which indicates there is no significant difference in variances between the groups. This means there is homogeneity of variance. Assumption 2 holds. Since the assumption 1 and 2 hold, we can say the model indicated the main effect of Region is significant.

References

Curly, T. (2017). Two-way ANOVA. Retrieved from: https://rpubs.com/tmcurley/twowayanova

Quick, J. (2011). R Tutorial Series: Two-Way ANOVA with Interactions and Simple Main Effects. Retrieved from: https://www.r-bloggers.com/r-tutorial-series-two-way-anova-with-interactions-and-simple-main-effects/

Quick, J. (2011). R Tutorial Series: Two-Way ANOVA with Unequal Sample Sizes. Retrieved from: https://www.r-bloggers.com/r-tutorial-series-two-way-anova-with-unequal-sample-sizes/

Ralph. (2010). Two-way Analysis of Variance (ANOVA). Retrieved from: https://www.r-bloggers.com/two-way-analysis-of-variance-anova/

R Statistics and Research. (2018). R-Two-Way ANOVA (part 1). Retrieved from: https://www.youtube.com/watch?v=yhHvzlJYqQY

R Statistics and Research. (2018). R-Two-Way ANOVA (part 2). Retrieved from:

part 2: https://www.youtube.com/watch?v=kr5K5-pBkXU&t=17s

Ralph. (2010). Two-way Analysis of Variance (ANOVA). Retrieved from: https://www.r-bloggers.com/two-way-analysis-of-variance-anova/

STHDA. (n.d.). Two-Way ANOVA Test in R. Retrieved from: http://www.sthda.com/english/wiki/two-way-anova-test-in-r