Region + Gender Two Way ANOVA

Sean O'Malley

4/16/2017

Discussion 1:

What questions can be answered by two-way ANOVA? Give examples.

- Two way ANOVA answers that compare three or more levels invovling two factors each with multiple levels. Essentially, there are two independent variables. Two examples of two-way ANOVA are:
    - An Engineer wants to assess the relationship between sintering time and the compressive strenght of three different metals. The engineer measures compressive strength of 5 specimens of each metal type at each sintering time: 100 mins, 150 mins, and 200 mins. To test the statitical significance of the two factors, and the interaction between them, the engineer uses two-way ANOVA.
    - Lets suppose that HR wants to know if occupational stress varies according to age and gender. The two independent variables are age and gender. Further, suppose the employees have been classified into thre groups: <40, 40-55, >55. Additionally, employees are listed as male and female. We can then have one observation per cell, one occupational stress score from one emloyee in each of the 6 cells. There are two hypothesis. H^1 All groups have equal stress. H^2 Both gender groups have equal stress on average. A two-way ANOVA then satisfies all 3 principles of design of experiments namely replication, randomization, and local control

What is an interaction effect?

- Interactions are when the effect of two, or more, variables is not simply additive. This page describes the interaction between two variables. It is possible to examine the interactions of three or more variables.

Should we include the main effects (regardless of the results of the test) in the model if the interaction term is statistically significant? Discuss.

- A main efect of an independent variable is the effect of the variable averaging over the levels of the other variables. A marginal mean for a level of variable is the mean of the means of all levels of the other variable. ANOVA tests main effects and interactions for significance, and we should consider this significance in our ANOVA output. Always keep in mind that the lack of evidence for an effect does not justify the conclusion that there is no effect. In other words, you do not accept the null hypothesis just because you do not reject it.

Discussion 2:

Given an interaction effect plot, how do you identify that there is an interaction effect in that plot?

- The effect of one factor depends on the level of the other factor. This is where we use an interaction plot to visualize possible interactions. Parallel lines indicate no interactions, and the greater the slope difference between the lines, the higher degree of interaction. Nevertheless, the interaction plot doesn't indicate statistical significance of this interaction.

What commands in R can be used to plot the interaction effect?

- 'interaction.plot(datAge, datDrug, dat$Value)'

List steps in two-way ANOVA using R

- Read in Data

- read.csv

- Fit the model using linear modeling and an interaction term, set to object results
  - results <- lm(Response ~ FactorA + FactorB + FactorA*FactorB, data = df)

- Plot interaction between factors
  - interaction.plot(factorA, factorB, Response)

- Fit a two way ANOVA with an interaction term
  - anova(results)

Two-Way ANOVA: Sex, Region and Income

Below we will fit regression models using regression methods of the syntax from multiple regression, but we will not allow the explanatory varaibles to be either categorical or quantitative.

Two-way ANOVA is used to compare the means of populations that are classified in two different ways, or the mean responses in an experiment with two factors.

Data Ingest

```
df <- read.csv("/Users/SeanOMalley1/Desktop/MSDS_660_Stats/twowayanova_data.csv")
```

EDA

```
summary(df)
```

```
##    Region       Male            Female      
##  East :3   Min.   : 40.00   Min.   : 45.00  
##  North:3   1st Qu.: 52.25   1st Qu.: 62.50  
##  South:3   Median : 71.00   Median : 76.50  
##  West :3   Mean   : 77.00   Mean   : 76.58  
##            3rd Qu.: 88.25   3rd Qu.: 89.00  
##            Max.   :150.00   Max.   :115.00
```

```
glimpse(df)
```

```
## Observations: 12
## Variables: 3
## $ Region <fctr> North, North, North, South, South, South, East, East, ...
## $ Male   <int> 50, 60, 45, 40, 53, 68, 92, 74, 87, 120, 150, 85
## $ Female <int> 45, 72, 65, 55, 75, 48, 87, 105, 79, 95, 78, 115
```

EDA: Data Manipulation using **tidyr**

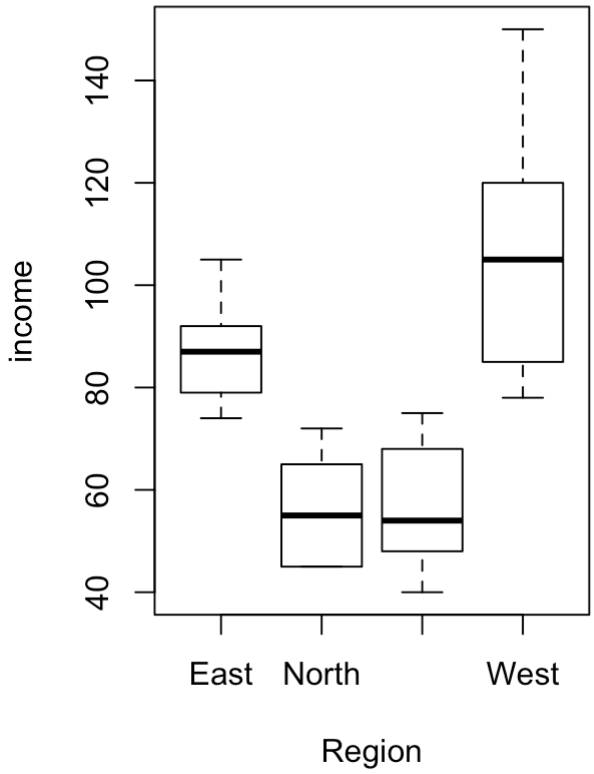
```
df <- gather(df, sex, income, -Region)

glimpse(df)
```

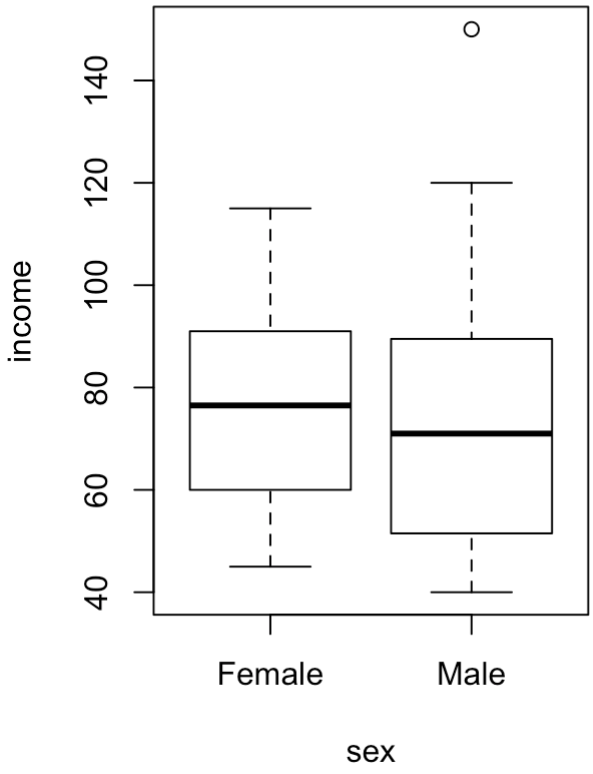
```
## Observations: 24
## Variables: 3
## $ Region <fctr> North, North, North, South, South, South, East, East, ...
## $ sex    <chr> "Male", "Male", "Male", "Male", "Male", "Male", "Male",...
## $ income <int> 50, 60, 45, 40, 53, 68, 92, 74, 87, 120, 150, 85, 45, 7...
```

```
df$sex <- as.factor(df$sex)
df$Region <- as.factor(df$Region)
df$income <- as.numeric(df$income)

summary(df)
```

```
##    Region      sex         income
##  East :6   Female:12   Min.   : 40.00
##  North:6   Male  :12   1st Qu.: 54.50
##  South:6               Median : 74.50
##  West :6               Mean   : 76.79
##                        3rd Qu.: 88.25
##                        Max.   :150.00
```

```
par(mfrow=c(1,2))

plot(income ~ Region + sex, data = df)
```





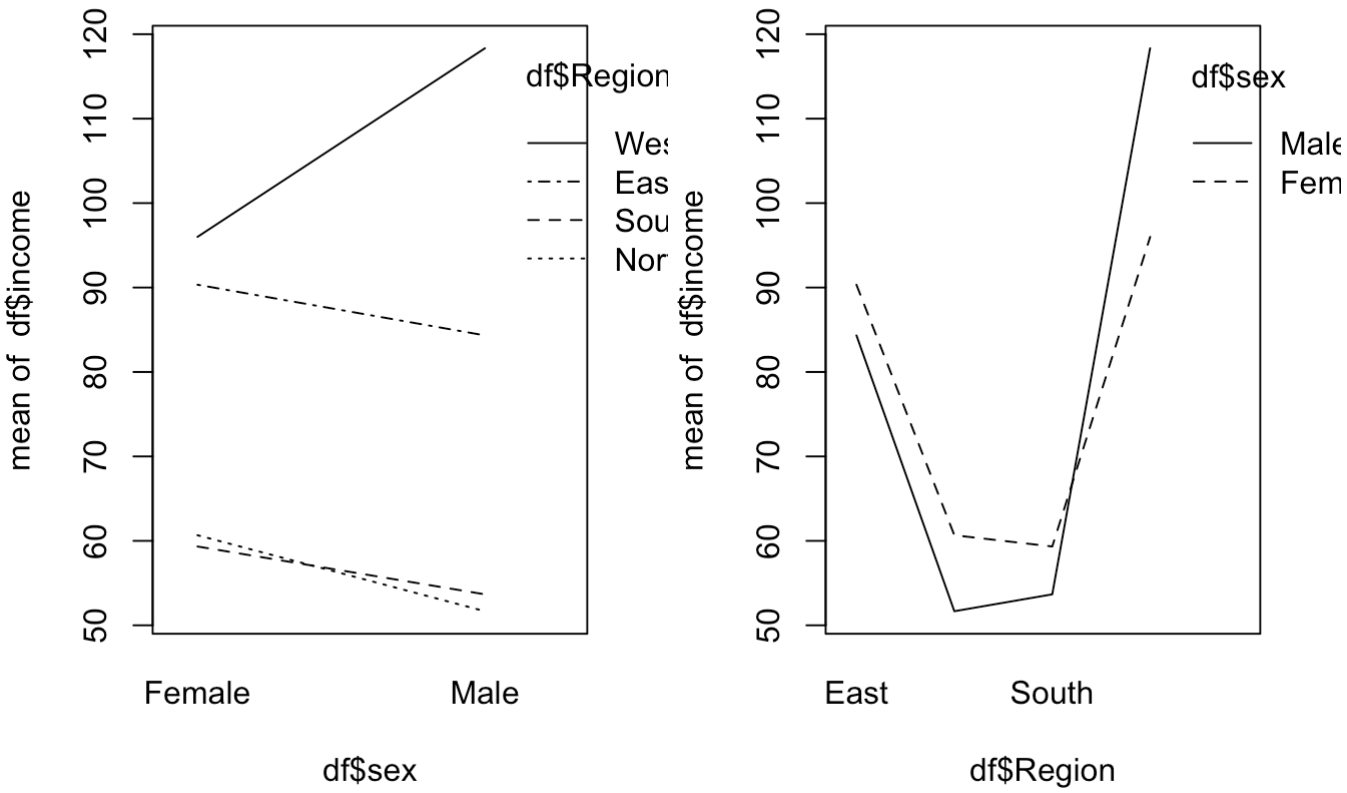
Hypothesis

H^0 : All regions and sexes have equal income on the average

H^1 : Region and sex have a significant effect on income on the average

H^2 : Region and sex are independent or that interaction effect is not present

Plot Interaction



With the naked eye test, it looks like sex has a noticeable interaction effect in the west and east regions, and only slight interaction visual effect in the south and north. When assessing both visualizations, regions appear to have a much larger on the interactions than the sex of the individual.

Two-Way ANOVA with Interactions

```
results <- lm(income ~ Region + sex + Region*sex, data = df)

anova(results)
```

```
## Analysis of Variance Table
##
## Response: income
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## Region      3 11225.5  3741.8 12.9456 0.0001509 ***
## sex         1     1.0     1.0  0.0036 0.9528734
## Region:sex  3   970.8   323.6  1.1196 0.3705135
## Residuals  16  4624.7   289.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

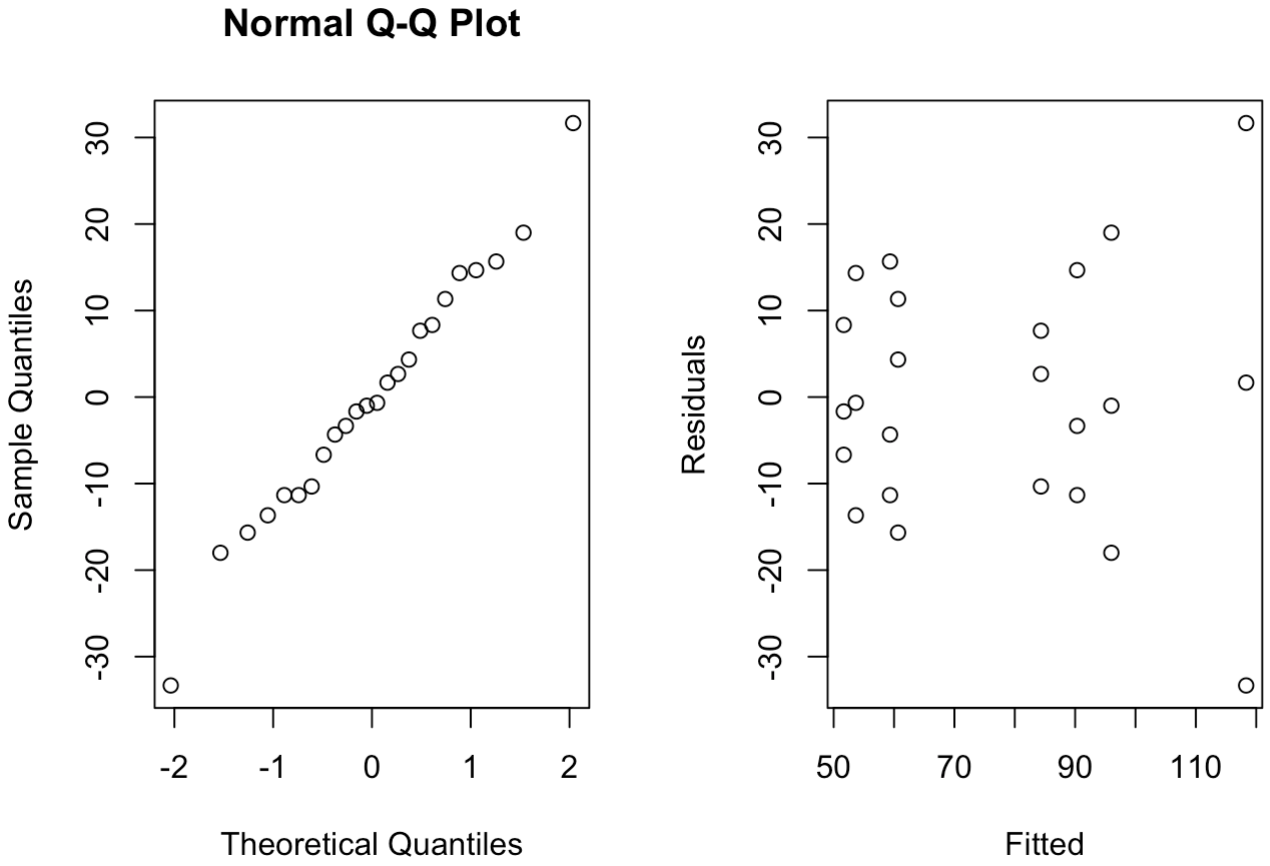
Interpreting Results

Looking at the ANOVA output, we see that there is no significant interaction effect with a p-value of 0.37 between region and sex. Region appears to have had a significant effect, with a p value around 0.00151 and and f value of 12.9. The sex however has an extremely low f value, confirming little evidence of the possibility of sex being a

factor. This is confirmed visually with the interaction effect in my opinion. Not that sex doesnt have any effect whatsoever, but region has markedly stark differences in income in comparison to sex of the person and the two factors really don’t seem to interact together, also confirmed in the statistical output above.

Check Assumptions

```
par(mfrow=c(1,2))

qqnorm(results$residuals)

plot(results$fitted.values, results$residuals, xlab = "Fitted", ylab = "Residuals")
```



Fitted values appear to be clustered more than normal vs the residuals, however in my opinion this can be attributed to the size of the data set and the stark significance of the regional income data. The residuals do not appear to violate an assumptions.

Lastly, H^0 and H^1 are rejected, the two factors do not have a significant interaction effect.