Lab 8: ANOVA in R

In this lab, we will discuss how to do ANOVAs in R. ANOVA stands for ANanlysis Of VAriance and examines whether the variance due to an independent variable or variables is significantly greater than the natural variance within a series of data. The idea is that if an independent variable can explain a significantly greater proprotion of variance than the natural variance in the data, then that variable is important.

The data for this assignment are a set of data examining how various factors predict a person's wage. You can find the data on Canvas as "Lab8Wage.csv".

These data are from 1985, from the Economics Web Institute surveying people on various attributes including how much they make. This may explain why the actual wages in this dataset are so small compared to waht would be expected today.

In this dataset, we are going to explore how regression can help us answer the question.

The variables are as follows:

- ID: person ID
- WAGE: wage (dollars per hour)
- OCCUPATION: occupation(1=Management, 2=Sales, 3=Clerical , 4=Service, 5=Professional, 6=Other)
- SECTOR: sector of employment(0= other, 1=Manufacturing, 2=Construction)
- UNION: Union membership (1=yes, 0=no)
- EDUCATION: Years of education (12 = high school diploma, 16= completed college, etc.)
- EXPERIENCE: Years of work experience
- AGE: Age in years
- SEX: Sex (0 – male, 1 – female)
- MARR: Married (0 – no, 1 – yes)
- RACE: Race (0 – other, 1 – white, 2 – Hispanic)
- SOUTH:Southern region (1 – yes, 0 – no)

Step 1: Create a folder for this assignment and download the data "Lab8Wage.csv" to that folder. Make a new R notebook and save it to that folder. Add a code chunk loading the data to the dataframe `d`.

When we have data that are coded as numbers, like here, we need to tell R that the numbers represent factors, or categorical data and that the numbers do not have numerical meaning. For example, the variable OCCUPATION has numbers for the categories. Management (1) is not less or more than Sales (2) and so forth.

Step 2: To do this, we have to use two commands. First is the `as.factor()` command, which turns a variable into a factor variable. To do this, type:

```
d$OCCUPATION = as.factor(d$OCCUPATION)
```

Step 3: Then we want to tell R what the levels mean, which makes it easier to interpret the output. We want to tell R that the levels represent different categories rather than numbers. Since the factors go in order from Management, Sales, etc., we can use the `levels()` command to change the order.

```
levels(d$OCCUPATION) = c('Management','Sales','Clerical','Service','Professional',
'Other')
```

Step 4: Use the `aggregate()` command to see what the mean and standard deviation of the WAGE variable is for each level of OCCUPATION. Note that OCCUPATION is the grouping variable and WAGE is the dependent variable.

Step 5: Now we will do the ANOVA. Type the following as a code chunk. The first line conducts the ANOVA and makes an object for it, and the second line outputs the ANOVA table:

```
x = aov(WAGE~OCCUPATION, data=d)
summary(x)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## OCCUPATION    5   2307   461.3   23.13 <2e-16 ***
## Residuals   527  10509    19.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The first row of this table corresponds to the Between Subjects variance, determined by the variable OCCUPATION. The second row is the "Residuals" which corresponds to the Within Subjects variance. The table provides degrees of freedom (Df), the Sums of Squares (Sum Sq), the Mean Squares (Mean Sq), and then the F-value and p-value (Pr(>F)). In this case the p-value is so small that R gives the scientific notation 2e-16, which is the smallest number that R can calcuate. As we mentioned in previous lectures, that means we just say $p < .001$.

To report these data in APA format, we would type: $F(5,527) = 23.13$, $p < .001$.

The next thing we are going to do is see which means are different using a post-hoc comparison, using the `pairwise.t.test()` command. As we mentioned in class, we can't compare each pair of groups using lots of t-tests, because that increases the likelihood of false positives, or Type 1 errores. The `pairwise.t.test()` command corrects for this by using a post-hoc correction. R's default is the 'holm' correction, which is fine for this class.

Step 6: Type the following to generate the pairwise t-test comparisons:

```
pairwise.t.test(d$WAGE,d$OCCUPATION)
```

When you do this, you have a table with rows and columns, and values in the middle. Each number represents a p-value for comparison between the row and column. For instance, the first row is Sales and the first Column is Management. The first value, in the top-left is the comparison between Sales and Management. The number here is 1.8e-05, which is scientific notation for .000018, which is well below .05. So the mean wages for Sales and Management are significantly different. If the number is 1.000, that means the difference is not significant at all.

Add in your annotations which other groups are significantly different from Management.

Repeating ANOVAs for SECTOR and RACE

Now we are going to do the same for sector and race.

Step 7 To convert those variables to factor variables, type the following:

```
d$SECTOR = as.factor(d$SECTOR)
levels(d$SECTOR) = c('Other','Manufacturing','Construction')
d$RACE = as.factor(d$RACE)
levels(d$RACE) = c('Other','White','Hispanic')
```

Step 8: Now go ahead and do an ANOVA for SECTOR as a factor and WAGE as the dependent variable. Report your F-value and p-value using APA format in your annotation and note whether there is a significant difference.

Step 9: Use `pairwise.t.test()` to do a post-hoc comparison for SECTOR to see which sectors are significantly different. Note which differences are significant in your annotation.

Step 10: Do Steps 8 and 9 for RACE, conducting the ANOVA and then the post-hoc comparison.

Step 11: Knit and submit.