

Intro to Data Science

Project 1: ANOVA analysis

Chunwang Yuan, Rajendra Kumar Vechalapu, Saketh Yelamarthi
Fall 2021

Outline

1. Abstract
2. Theory
3. Exploratory Data analysis
4. Analysis Results & Explanation
5. Conclusion

1 Abstract

After detecting that a factor has a significant effect on a dependent variable globally, we often want to go further into detail and ask which specific factor levels differ from one another. Running a one-way ANOVA on the data would answer the very general question: “is there at least one treatment which significantly differs from the others?” If the ANOVA test is significant, another question could be asked: “what treatments differ from one another?” This question requires testing the differences between all pairs of treatments. Pairwise multiple comparisons tools were developed to address this issue.

2 Theory

Type *Markdown* and LaTeX: α^2

3 Exploratory Data analysis

In this section, we use the method of One-way Anova to explore the data

In [2]:

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
from statsmodels.formula.api import ols
```

Firstly, we load the data as the dataframe type set.

In [3]:

```
data = pd.read_csv('DSCI6002_prj1_data.csv', header = None)
data_array = np.array(data)
df = pd.DataFrame(data_array ,columns = ['treatments','value'])
df.dtypes
```

Out[3]:

```
treatments    object
value         object
dtype: object
```

Since the data type of the column of 'value' is object, we need to convert the value from object to numeric.

In [4]:

```
# convert the value column to numeric type
df['value'] = pd.to_numeric(df['value'])
df.dtypes
```

Out[4]:

```
treatments    object
value         float64
dtype: object
```

In [5]:

```
model = ols('value~C(treatments)', data = df ).fit()
model.summary()
```

Out[5]:

OLS Regression Results

| | | | |
|--------------------------|------------------|----------------------------|---------|
| Dep. Variable: | value | R-squared: | 0.016 |
| Model: | OLS | Adj. R-squared: | 0.012 |
| Method: | Least Squares | F-statistic: | 4.619 |
| Date: | Thu, 18 Nov 2021 | Prob (F-statistic): | 0.00106 |
| Time: | 15:53:06 | Log-Likelihood: | -4837.1 |
| No. Observations: | 1172 | AIC: | 9684. |
| Df Residuals: | 1167 | BIC: | 9709. |
| Df Model: | 4 | | |
| Covariance Type: | nonrobust | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | </ | |

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

by implementing the method of One-way Anova to the model, we can get the parameters.

In [6]:

```
### One-way ANOVA
anova_table = sm.stats.anova_lm(model, typ = 2)
anova_table
```

Out[6]:

| | sum_sq | df | F | PR(>F) |
|----------------------|---------------|--------|----------|----------|
| C(treatments) | 4176.452688 | 4.0 | 4.618904 | 0.001056 |
| Residual | 263802.838942 | 1167.0 | NaN | NaN |

In [7]:

```
esq_sm = anova_table['sum_sq'][0]/(anova_table['sum_sq'][0]+anova_table['sum_sq'][1])
anova_table['EtaSq'] = [esq_sm, 'NaN']
anova_table
```

Out[7]:

| | sum_sq | df | F | PR(>F) | EtaSq |
|----------------------|---------------|--------|----------|----------|----------|
| C(treatments) | 4176.452688 | 4.0 | 4.618904 | 0.001056 | 0.015585 |
| Residual | 263802.838942 | 1167.0 | NaN | NaN | NaN |

By applying the method of Multiple Pairwise Comparisons and Turkey HSD test, the results show us that the difference between each group.

In [8]:

```
pair_t = model.t_test_pairwise('C(treatments)')
pair_t.result_frame
```

Out[8]:

| | coef | std err | t | P> t | Conf. Int. Low | Conf. Int. Upp. | pvalue- hs | reject- hs |
|--------------------------------|-----------|----------|-----------|----------|-------------------|--------------------|---------------|---------------|
| Graduate-Bachelor's | 1.496838 | 1.533586 | 0.976038 | 0.329248 | -1.512057 | 4.505733 | 0.698223 | False |
| HS-Bachelor's | -2.097484 | 1.143460 | -1.834331 | 0.066859 | -4.340952 | 0.145984 | 0.292485 | False |
| Jr Coll-Bachelor's | -1.006255 | 1.795528 | -0.560423 | 0.575299 | -4.529079 | 2.516569 | 0.760113 | False |
| Less than HS-Bachelor's | -5.566798 | 1.661832 | -3.349796 | 0.000835 | -8.827311 | -2.306286 | 0.007486 | True |
| HS-Graduate | -3.594322 | 1.368362 | -2.626733 | 0.008734 | -6.279048 | -0.909597 | 0.067772 | False |
| Jr Coll-Graduate | -2.503093 | 1.946493 | -1.285950 | 0.198716 | -6.322110 | 1.315924 | 0.587763 | False |
| Less than HS-Graduate | -7.063636 | 1.823897 | -3.872826 | 0.000114 | -10.642120 | -3.485153 | 0.001135 | True |
| Jr Coll-HS | 1.091230 | 1.656638 | 0.658701 | 0.510218 | -2.159092 | 4.341551 | 0.760113 | False |
| Less than HS-HS | -3.469314 | 1.510700 | -2.296494 | 0.021824 | -6.433306 | -0.505322 | 0.143123 | False |
| Less than HS-Jr Coll | -4.560544 | 2.049057 | -2.225679 | 0.026226 | -8.580791 | -0.540296 | 0.147393 | False |

In [9]:

```
mc = sm.stats.multicomp.MultiComparison(df['value'], df['treatments'])
mc_results = mc.tukeyhsd()
print(mc_results)
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05

| group1 | group2 | meandiff | p-adj | lower | upper | reject |
|------------|--------------|----------|--------|----------|---------|--------|
| Bachelor's | Graduate | 1.4968 | 0.8514 | -2.693 | 5.6867 | False |
| Bachelor's | HS | -2.0975 | 0.3547 | -5.2215 | 1.0265 | False |
| Bachelor's | Jr Coll | -1.0063 | 0.9 | -5.9117 | 3.8992 | False |
| Bachelor's | Less than HS | -5.5668 | 0.0074 | -10.107 | -1.0266 | True |
| Graduate | HS | -3.5943 | 0.0664 | -7.3327 | 0.1441 | False |
| Graduate | Jr Coll | -2.5031 | 0.6756 | -7.821 | 2.8148 | False |
| Graduate | Less than HS | -7.0636 | 0.0011 | -12.0466 | -2.0807 | True |
| HS | Jr Coll | 1.0912 | 0.9 | -3.4348 | 5.6172 | False |
| HS | Less than HS | -3.4693 | 0.1466 | -7.5966 | 0.658 | False |
| Jr Coll | Less than HS | -4.5605 | 0.1711 | -10.1587 | 1.0376 | False |

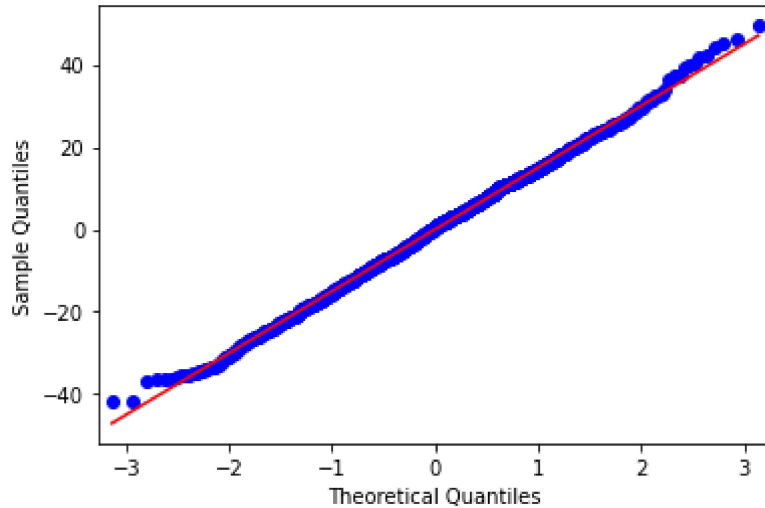
One-way ANOVA: Assumptions checks

In [10]:

```
res = model.resid
```

In [11]:

```
fig = sm.qqplot(res, line = 's')
```



In [12]:

```
import seaborn as sns
```

In [13]:

```
### Normality Assumption check
from scipy import stats

#using the Shapiro-wilk test
w_shapiro, pvalue_shapiro = stats.shapiro(res)
print(pvalue_shapiro)
```

```
0.33162370324134827
```

In [15]:

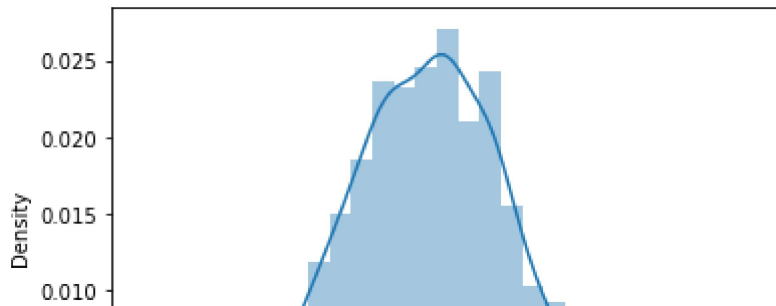
```
sns.distplot(res,bins = 'auto',hist = True)
```

D:\program\Anaconda\lib\site-packages\seaborn\distributions.py:2551: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).

```
warnings.warn(msg, FutureWarning)
```

Out[15]:

```
<AxesSubplot:ylabel='Density'>
```



In [26]:

```
# Using Bartlett's test
w_bartlett, pvalue_bartlett = stats.bartlett(df['value'][df['treatments'] == 'Less than HS'],
                                             df['value'][df['treatments'] == 'HS'],
                                             df['value'][df['treatments'] == 'Jr Coll'],
                                             df['value'][df['treatments'] == "Bachelor's"],
                                             df['value'][df['treatments'] == "Graduate"])
print("Bartlett's test:\tw:{:7.4f}, pvalue:{:7.4f}".format(w_bartlett, pvalue_bartlett))
```

```
Bartlett's test:      w:17.7424, pvalue: 0.0014
```

In [25]:

```
#Using Levene Variance test
w_levene, pvalue_levene = stats.levene(df['value'][df['treatments'] == 'Less than HS'],
                                       df['value'][df['treatments'] == 'HS'],
                                       df['value'][df['treatments'] == 'Jr Coll'],
                                       df['value'][df['treatments'] == "Bachelor's"],
                                       df['value'][df['treatments'] == "Graduate"])
print("Levene's test:\tw:{:7.4f}, pvalue:{:7.4f}".format(w_levene, pvalue_levene))
```

```
Levene's test:  w: 5.7617, pvalue: 0.0001
```

4 Analysis Results & Explanation

In [27]:

```
anova_table
```

Out[27]:

| | sum_sq | df | F | PR(>F) | EtaSq |
|---------------|---------------|--------|----------|----------|----------|
| C(treatments) | 4176.452688 | 4.0 | 4.618904 | 0.001056 | 0.015585 |
| Residual | 263802.838942 | 1167.0 | NaN | NaN | NaN |

As the result of One-way Anova shows, there is at least one treatment which significantly differs from the others. So we are eager to know which one group is different from other group. By implying the method of Multiple Pairwise Comparisons and Turkey HSD test, we can clearly to check the specifical group that differs from the others.

In [13]:

```
print(mc_results)
pair_t.result_frame
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05

```
=====
group1      group2      meandiff p-adj      lower      upper      reject
-----
Bachelor's   Graduate      1.4968 0.8514     -2.693     5.6867     False
Bachelor's   HS            -2.0975 0.3547     -5.2215     1.0265     False
Bachelor's   Jr Coll      -1.0063 0.9         -5.9117     3.8992     False
Bachelor's   Less than HS -5.5668 0.0074     -10.107     -1.0266     True
Graduate     HS            -3.5943 0.0664     -7.3327     0.1441     False
Graduate     Jr Coll      -2.5031 0.6756     -7.821      2.8148     False
Graduate     Less than HS -7.0636 0.0011     -12.0466     -2.0807     True
HS           Jr Coll      1.0912 0.9         -3.4348     5.6172     False
HS           Less than HS -3.4693 0.1466     -7.5966     0.658      False
Jr Coll      Less than HS -4.5605 0.1711     -10.1587     1.0376     False
=====
```

Out[13]:

| | coef | std err | t | P> t | Conf. Int. Low | Conf. Int. Upp. | pvalue- hs | reject- hs |
|--------------------------------|-----------|----------|-----------|----------|-------------------|--------------------|---------------|---------------|
| Graduate-Bachelor's | 1.496838 | 1.533586 | 0.976038 | 0.329248 | -1.512057 | 4.505733 | 0.698223 | False |
| HS-Bachelor's | -2.097484 | 1.143460 | -1.834331 | 0.066859 | -4.340952 | 0.145984 | 0.292485 | False |
| Jr Coll-Bachelor's | -1.006255 | 1.795528 | -0.560423 | 0.575299 | -4.529079 | 2.516569 | 0.760113 | False |
| Less than HS-Bachelor's | -5.566798 | 1.661832 | -3.349796 | 0.000835 | -8.827311 | -2.306286 | 0.007486 | True |
| HS-Graduate | -3.594322 | 1.368362 | -2.626733 | 0.008734 | -6.279048 | -0.909597 | 0.067772 | False |
| Jr Coll-Graduate | -2.503093 | 1.946493 | -1.285950 | 0.198716 | -6.322110 | 1.315924 | 0.587763 | False |
| Less than HS-Graduate | -7.063636 | 1.823897 | -3.872826 | 0.000114 | -10.642120 | -3.485153 | 0.001135 | True |
| Jr Coll-HS | 1.091230 | 1.656638 | 0.658701 | 0.510218 | -2.159092 | 4.341551 | 0.760113 | False |
| Less than HS-HS | -3.469314 | 1.510700 | -2.296494 | 0.021824 | -6.433306 | -0.505322 | 0.143123 | False |
| Less than HS-Jr Coll | -4.560544 | 2.049057 | -2.225679 | 0.026226 | -8.580791 | -0.540296 | 0.147393 | False |

According to the results of Multiple Pairwise Comparisons and Turkey HSD test above, the reject status between group 'Less than HS' and "Bachelor's" is True as the same as between group 'Less than HS' and 'Graduate', and rest of groups are False. That illustrates the group of 'Less than HS' is significantly different from groups of 'Bachelor's' and 'Graduate'.

In [30]:

```
print("Shapiro-wilk's test:\tw:{:7.4f}, pvalue:{:7.4f}".format(w_shapiro, pvalue_shapiro))
print("Bartlett's test:\tw:{:7.4f}, pvalue:{:7.4f}".format(w_bartlett, pvalue_bartlett))
print("Levene's test:\tw:{:7.4f}, pvalue:{:7.4f}".format(w_levene, pvalue_levene))
```

```
Shapiro-wilk's test:    w: 0.9984, pvalue: 0.3316
Bartlett's test:       w:17.7424, pvalue: 0.0014
Levene's test:    w: 5.7617, pvalue: 0.0001
```

For checking the normality, we use the method of Shapiro-wilk test.

Under the method of Shapiro-wilk test, the null-hypothesis of this test is that the population is normally distributed. The p value is greater than the chosen alpha level, then the null hypothesis (that the data came from a normally distributed population) can not be rejected (e.g., for an alpha level of $\alpha = 0.05$).

For identifying whether there are equal variances between groups, we imply the methods of Bartlett's test and Levene Variance test.

We are testing the null hypothesis that the batch variances are all equal. Because the pvalue is less than the alpha value, we reject the null hypotheses at the 0.05 significance level and conclude that at least one batch variance is different from the others.

5 Conclusion

According to the results of one-way ANOVA and Multiple Pairwise Comparisons and Turkey HSD test, we can see the group of 'Less than HS' has a significant difference with the two groups of "Bachelor's" and 'Graduate'.

Also, the data is a normal distribution.

In []:

In []: