

Unit 14: Analysis of Variance (ANOVA)

Case Study: Analysis of the Relationship between Age and Political Affiliation

A very common special case of linear regression models is when there is one categorical explanatory variable, and the goal is to determine if the **mean response** is significantly different between categories of the explanatory variable.

This is an extension of two-sample analysis where we have one categorical explanatory variable, such as website version A versus B, and we compare them based on a quantified outcome. The ANOVA approach lets us compare $K > 2$ categories at the same time.

```
In [1]: import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

import statsmodels.api as sm
import statsmodels.formula.api as smf
```

1. Analyses for Associations

See Unit 14 slides (section 1)

2. Association Analyses Summary: One Categorical Explanatory Variable (with >2 Levels)-> Numerical Response Variable (*One Way*)

See Unit 14 slides (section 2)

3. Association Analyses Summary: One Categorical Explanatory Variable (with >2 Levels)-> Numerical Response Variable (*Another Way*)

See Unit 14 slides (section 3)

4. Modeling the Association Between a Categorical Explanatory Variable (with >2 levels) and a Numerical Response Variable (in the Sample).

See Unit 14 slides (section 4)

Let's again examine our random sample of adults living in the U.S. (from 2017) from Pew Research.

```
In [2]: missing_values = ["NaN", "nan", "Don't know/Refused (VOL.)"]
pew = pd.read_csv('Feb17public.csv',
                  na_values=missing_values)[['age', 'party']].dropna()
pew.head()
```

Out[2]:

	age	party
0	80.0	Independent
1	70.0	Democrat
2	69.0	Independent
3	50.0	Republican
4	70.0	Democrat

```
In [3]: pew.shape
```

Out[3]: (1465, 2)

How many of each political affiliation are there in this sample?

```
In [4]: pew['party'].value_counts()
```

```
Out[4]: Democrat      527  
Independent    525  
Republican      367  
No preference (VOL.)  41  
Other party (VOL.)   5  
Name: party, dtype: int64
```

Let's rename the party categories so they are easier to label in graphs. We can do this as follows.

```
In [5]: # rename categories so they display better  
party = pd.Categorical(pew['party'])  
party.rename_categories({'Democrat': 'Dem',  
                        'Independent': 'Ind',  
                        'Republican': 'Rep',  
                        'No preference (VOL.)': 'No_Pref',  
                        'Other party (VOL.)': 'Other'}, inplace=True)  
  
pew['party']=party
```

```
In [6]: pew['party'].value_counts()
```

```
Out[6]: Dem      527  
Ind      525  
Rep      367  
No_Pref    41  
Other      5  
Name: party, dtype: int64
```

Next, let's fit our multiple linear regression model for the sample, where age is our response variable and party is our explanatory variable.

```
In [7]: agemod = smf.ols('age ~ party', data=pew).fit()
agemod.summary()
```

Out[7]: OLS Regression Results

Dep. Variable:	age	R-squared:	0.052
Model:	OLS	Adj. R-squared:	0.049
Method:	Least Squares	F-statistic:	19.82
Date:	Mon, 05 Apr 2021	Prob (F-statistic):	6.66e-16
Time:	17:29:42	Log-Likelihood:	-6261.1
No. Observations:	1465	AIC:	1.253e+04
Df Residuals:	1460	BIC:	1.256e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Intercept	50.4991	0.758	66.618	0.000	49.012	51.986
party[T.Ind]	-3.6914	1.073	-3.440	0.001	-5.796	-1.587
party[T.No_Pref]	-7.3527	2.821	-2.606	0.009	-12.887	-1.818
party[T.Other]	-5.8991	7.819	-0.754	0.451	-21.237	9.439
party[T.Rep]	6.2775	1.183	5.306	0.000	3.957	8.598

Omnibus:	130.613	Durbin-Watson:	1.725
Prob(Omnibus):	0.000	Jarque-Bera (JB):	40.798
Skew:	-0.017	Prob(JB):	1.38e-09
Kurtosis:	2.183	Cond. No.	19.0

Notes:

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

4.1. RELATIONSHIP BETWEEN THE SAMPLE INTERCEPT AND SLOPES OF THIS MULTIPLE LINEAR REGRESSION AND THE SAMPLE MEANS

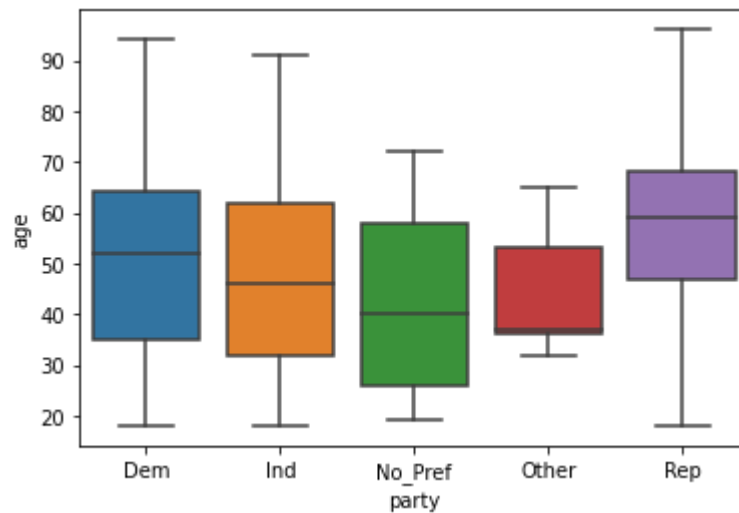
See Unit 14 (Section 4.1) slides.

Is there an association between party and age *in this sample*? Let's next use descriptive analytics to find out.

Visualization

With several groups we can use side by side boxplots to visualize the age distributions.

```
In [8]: sns.boxplot(x='party', y='age', data=pew)
plt.show()
```



The median age across the five different political affiliation groups is different, so we can say that there is at least some association *in the sample* between political affiliation and age.

Summary statistics

Using Pandas **groupby()** function to get summary statistics for each political affiliation

```
In [9]: # within group means
pew.groupby('party').mean()
```

Out[9]:

	age
party	
Dem	50.499051
Ind	46.807619
No_Pref	43.146341
Other	44.600000
Rep	56.776567

```
In [10]: # within group sample standard deviations
pew.groupby('party').std()
```

```
Out[10]:
```

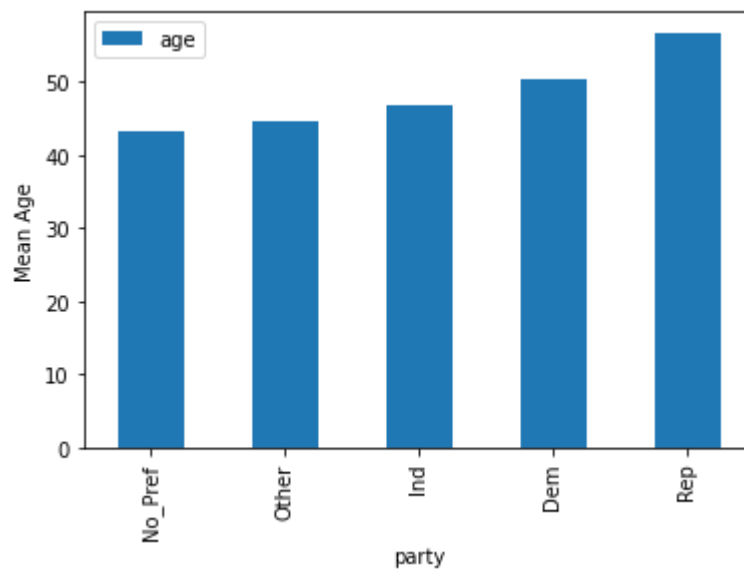
	age
party	
Dem	17.687279
Ind	17.517144
No_Pref	17.062475
Other	13.939153
Rep	16.885801

```
In [11]: # within group sample sizes
pew.groupby('party').count()
```

```
Out[11]:
```

	age
party	
Dem	527
Ind	525
No_Pref	41
Other	5
Rep	367

```
In [12]: pew.groupby('party').mean().sort_values(by='age').plot.bar()
plt.ylabel('Mean Age')
plt.show()
```



5. Modeling the Association Between a Categorical Explanatory Variable (with >2 levels) and a Numerical Response Variable (*in the Population*).

See Unit 14 slides (section 5)

Now let's think about what the multiple linear regression equation would look like if we modelled the relationship between age (response variable) and political affiliation (explanatory variable) in the population of all adults living in the U.S.

$$\hat{age} = \beta_0 + \beta_1 party[T. Ind] + \beta_2 party[T. Ind] + \beta_3 party[T. Nopref] + \beta_4 party[T. Rep]$$

Do we have sufficient evidence to suggest that at least one of the four population slopes is non-zero?

1. First let's formulate the hypotheses.

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

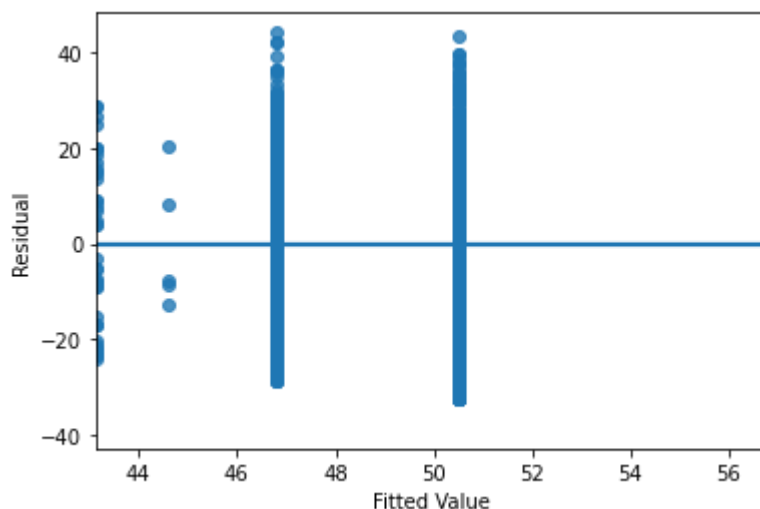
$$H_A : \text{At least one } \beta_i \neq 0 \text{ (for } i=1,2,3,4)$$

2. Next, let's check the conditions for conducting inference.

a.) Linearity Condition

It appears that there is an even distribution of points above and below the line in the fitted values vs. residuals plot as we move from left to right. So we can say that this condition is met.

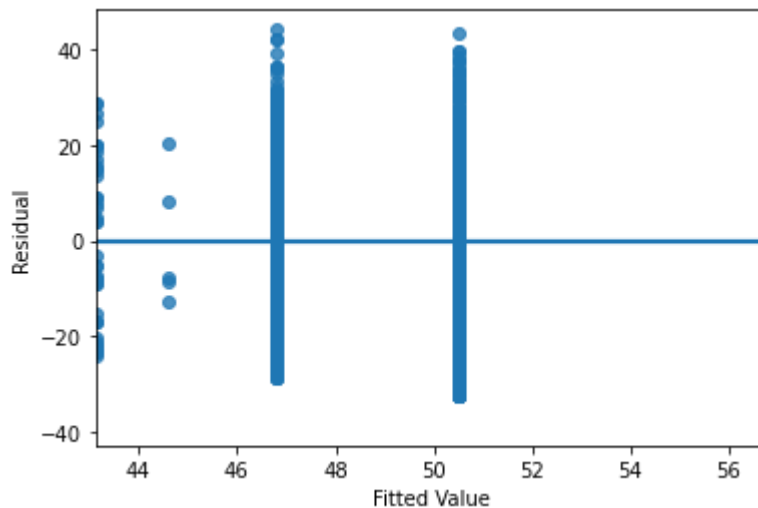
```
In [13]: sns.regplot(x=agemod.fittedvalues, y=agemod.resid, ci=None)
plt.ylabel('Residual')
plt.xlabel('Fitted Value')
plt.show()
```



b.) Constant Variance of Residuals Condition

It appears that the spread of the residuals (ie. the y-axis spread) in the fitted values vs. residuals plot changes as we move from left to right. So we cannot say that this condition is met.

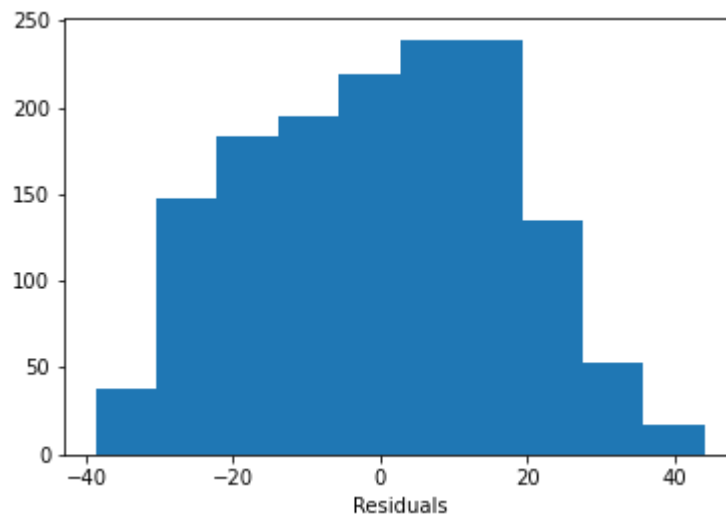
```
In [14]: sns.regplot(x=agemod.fittedvalues, y=agemod.resid, ci=None)
plt.ylabel('Residual')
plt.xlabel('Fitted Value')
plt.show()
```



c.) Normality of Residuals (with Mean 0) Condition

It appears that the residuals are slightly skewed to the right. So because the histogram of residuals is not quite symmetric and unimodal, this condition is somewhat not met.

```
In [15]: plt.hist(agemod.resid)
plt.xlabel('Residuals')
plt.show()
```



d.) Independence of Residuals Condition

While we cannot know for sure if this condition is met (using the methods that we know so far in this class), we know that this condition will NOT be met if the sample is not random or $n \geq 10\%$ of the population.

But, we do know that:

- this sample is random and
- $n = 1465 < 10\%$ of all adults living in the U.S..

So this tells us that this condition is not violated *in this particular way*.

e.) No Multicollinearity Condition

This linear regression model involves only one explanatory variable and it is categorical, so we do not need to check this condition.

So not all of the conditions for conducting inference on a population slope or intercept are quite met.

Thus some of the conclusions that we make about our hypotheses may be slightly off, but we will proceed with caution with that in mind.

3. Next, let's find the test statistic and the p-value that correspond to these hypotheses.

```
In [16]: ageomod.summary()
```

Out[16]: OLS Regression Results

Dep. Variable:	age	R-squared:	0.052
Model:	OLS	Adj. R-squared:	0.049
Method:	Least Squares	F-statistic:	19.82
Date:	Mon, 05 Apr 2021	Prob (F-statistic):	6.66e-16
Time:	17:29:43	Log-Likelihood:	-6261.1
No. Observations:	1465	AIC:	1.253e+04
Df Residuals:	1460	BIC:	1.256e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Intercept	50.4991	0.758	66.618	0.000	49.012	51.986
party[T.Ind]	-3.6914	1.073	-3.440	0.001	-5.796	-1.587
party[T.No_Pref]	-7.3527	2.821	-2.606	0.009	-12.887	-1.818
party[T.Other]	-5.8991	7.819	-0.754	0.451	-21.237	9.439
party[T.Rep]	6.2775	1.183	5.306	0.000	3.957	8.598
Omnibus:	130.613	Durbin-Watson:	1.725			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	40.798			
Skew:	-0.017	Prob(JB):	1.38e-09			
Kurtosis:	2.183	Cond. No.	19.0			

Notes:

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

The test statistic for this test is 19.82 and the p-value for this test is 6.66×10^{-16} .

4. Finally, use the p-value (and a significance level of $\alpha=0.05$) to make a conclusion about your hypotheses.

Because $p - value = 6.66 \times 10^{-16} \leq \alpha 0,05$ we reject the null hypothesis. Thus there is sufficient evidence to suggest that at least one of the population slopes in this model is non-zero.

5.1 ANOVA: RELATIONSHIP BETWEEN THE POPULATION SLOPES OF THIS MULTIPLE LINEAR REGRESSION AND THE POPULATION MEANS

See Unit 14 slides (section 5.1).

Do we have sufficient evidence to suggest that at least one pair of political affiliations (Democrat, Republican, Independent, No preference, Other party) have average ages out of all adults living in the U.S. that are different?

1. First let's formulate the hypotheses.

$$H_0 : \mu_{Dem} = \mu_{Ind} = \mu_{Other} = \mu_{Nopref} = \mu_{Rep}$$

H_A : At least one pair of groups has population mean values are different from each other.

2. Next, let's check the conditions for conducting inference (on the intercept and slopes $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ of the population model

$$\hat{age} = \beta_0 + \beta_1 party[T. Ind] + \beta_2 party[T. Ind] + \beta_3 party[T. Nopref] + \beta_4 party[T. Rep])$$

3. Next, let's find the test statistic and the p-value that correspond to these hypotheses?

```
In [17]: agemod = smf.ols('age ~ party', data=pew).fit()
agemod.summary()
```

Out[17]: OLS Regression Results

Dep. Variable:	age	R-squared:	0.052
Model:	OLS	Adj. R-squared:	0.049
Method:	Least Squares	F-statistic:	19.82
Date:	Mon, 05 Apr 2021	Prob (F-statistic):	6.66e-16
Time:	17:29:43	Log-Likelihood:	-6261.1
No. Observations:	1465	AIC:	1.253e+04
Df Residuals:	1460	BIC:	1.256e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Intercept	50.4991	0.758	66.618	0.000	49.012	51.986
party[T.Ind]	-3.6914	1.073	-3.440	0.001	-5.796	-1.587
party[T.No_Pref]	-7.3527	2.821	-2.606	0.009	-12.887	-1.818
party[T.Other]	-5.8991	7.819	-0.754	0.451	-21.237	9.439
party[T.Rep]	6.2775	1.183	5.306	0.000	3.957	8.598

Omnibus:	130.613	Durbin-Watson:	1.725
Prob(Omnibus):	0.000	Jarque-Bera (JB):	40.798
Skew:	-0.017	Prob(JB):	1.38e-09
Kurtosis:	2.183	Cond. No.	19.0

Notes:

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

The test statistic for this test is 19.82 and the p-value for this test is 6.66×10^{-16} .

4. Finally, use the p-value (and a significance level of $\alpha=0.05$) to make a conclusion about your hypotheses.

Because $p - value = 6.66 \times 10^{-16} \leq \alpha 0,05$ we reject the null hypothesis. Thus there is sufficient evidence to at least one pair of population mean ages (out of the five political affiliation groups) are not equal to each other.

In []: