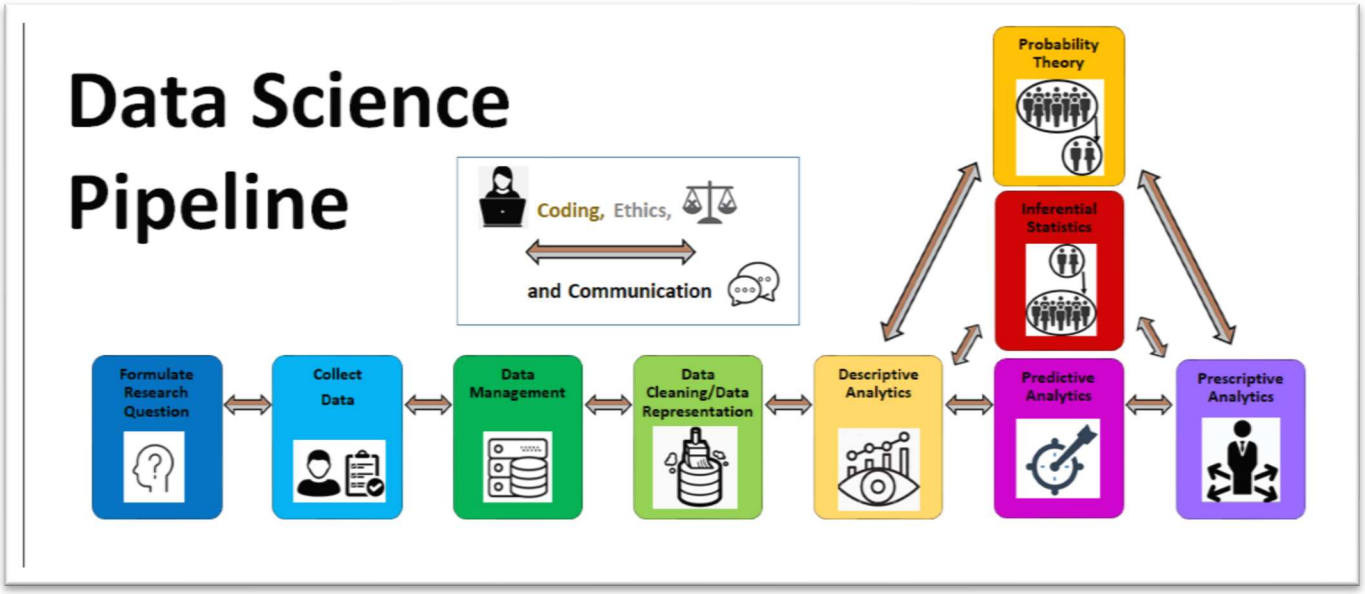
Unit 14: Analysis of Variance (ANOVA)

Case Studies:

- Is there an association between **political affiliation** (republican/democrat/independent/"no preference"/"other party") and **age**?

Purpose of this Lectures



1. **Analyses for Associations**
2. **Association Analyses Summary:** One Categorical Explanatory Variable (with >2 Levels)-> Numerical Response Variable (One Way)
3. **Association Analyses Summary:** One Categorical Explanatory Variable (with >2 Levels)-> Numerical Response Variable (Another Way)
4. **Modeling the Association** Between a **Categorical Explanatory Variable (with >2 levels)** and **a Numerical Response Variable** (in the Sample).
   4.1. Relationship between the sample intercept and slopes and the sample means.
5. **Modeling the Association** Between a **Categorical Explanatory Variable (with >2 levels)** and **a Numerical Response Variable** (in the Population).
   5.1. **ANOVA**: Relationship between the population intercept and slopes and the population means

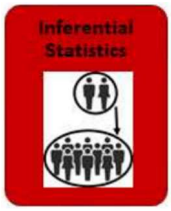
Additional Resources

Section 5.5 in Diez, Barr, and Cetinkaya-Rundel, (2015), *OpenIntro Statistics* https://www.openintro.org/download.php?file=os3&redirect=/stat/textbook/os3.php

1. Analyses for Associations

Questions to consider, when selecting an analysis to test an association.

1. Which variable is the **response variable** in this association?
    - **a.** Is it a **categorical** or **numerical** variable?
    - **b.** If it's a categorical variable, **how many levels** does it have?
2. Which variable(s) is the **explanatory variable** in this association?
    - **a.** Is it a **categorical** or **numerical** variable?
    - **b.** If it's a categorical variable, **how many levels** does it have?
3. How would you **quantify this association**?
    - **a.** Difference between two summary statistics? What two summary statistics?
    - **b.** With a model? What kind of model?



4. Are you interested in an association in a **sample** or a **population**?
5. When is it **appropriate to use this test** for association?

6. Can you use this model/test to **make predictions**?
    - **a.** How would you quantify the performance of your predictions?

2. ASSOCIATION ANALYSIS SUMMARY: *(ONE WAY TO PERFORM THIS ANALYSIS)*

RESPONSE: NUMERICAL

EXPLANATORY: ONE CATEGORICAL *WITH >2 LEVELS*

Research Questions about Associations	**Type of Variables Involved in the Association Test**	**Explanatory Variables:** One Categorical Variable *with >2 levels* **Response Variable:** Numerical Variable
	Example	Is there an association between **political affiliation** (republican/democrat/independent/"no preference"/"other party") and **age**? 
	Type of Association *(Way to Quantify Association)*	**Multiple Linear Regression Model** *(linear relationship between categorical explanatory variable (x) and response variable (y))*
Descriptive Analytics	**How to Describe an Association in a Sample?**	1. Multiple Linear Regression Model: • $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}x_1 + \widehat{\beta_2}x_2. + \cdots + \widehat{\beta_p}x_p$ • $R^2$of the model
	When is this analysis *(for the sample)* appropriate to use?	Linearity condition is met
Inferential Statistics	**How to Infer an Association for a Population?**	Conduct inference on: • A single population parameter β_i • All population parameters $\beta_1, \ldots, \beta_p$ • A subset of population parameters $\beta_{p+1}, \ldots, \beta_{p+q}$
	When is this analysis *(for the population)* appropriate to use?	1. Linearity condition is met 2. Constant variance of residuals condition is met. 3. Residuals are normal (and centered at 0). 4. Residuals are independent. 5. No-Multicollinearity condition is met.
Predictive Analytics	**Making Predictions**	Use your multiple linear regression line to make predictions $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}x_1 + \widehat{\beta_2}x_2. + \cdots + \widehat{\beta_p}x_p$
	How to quantify the performance of your prediction(s)?	• Individual Data Point: residual • All Data: root mean square error (RMSE)

3. ASSOCIATION ANALYSIS SUMMARY: *(ANOTHER WAY TO PERFORM THIS ANALYSIS)*

RESPONSE: NUMERICAL

EXPLANATORY: ONE CATEGORICAL *WITH >2 LEVELS*

Research Questions about Associations	**Type of Variables Involved in the Association Test**	**Explanatory Variables:** One Categorical Variable *with >2 levels* **Response Variable:** Numerical Variable
	Example	Is there an association between **political affiliation** (republican/democrat/independent/"no preference"/"other party") and **age**? 
	Type of Association *(Way to Quantify Association)*	*Does at least one pair of political affiliations (out of the five given) have mean ages that are different?*
Descriptive Analytics	**How to Describe an Association in a Sample?**	$\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p$ at least one pair of these sample means are different. • p is the number of levels of the categorical variable • $\bar{x}_i$ is the sample mean (the numerical variable) for the *ith* level of the categorical variable.
	When is this analysis *(for the sample)* appropriate to use?	Sample distribution *i* symmetric and unimodal (*for i=1,2,…,p).*
Inferential Statistics	**How to Infer an Association for a Population?**	**Conduct an Analysis of Variance (ANOVA)** $H_0: \mu_1 = \mu_2 = \cdots = \mu_p$ $H_A: \mu_i \neq \mu_j$ *for at least one pair of levels* • p is the number of levels of the categorical variable • $\bar{x}_i$ is the sample mean (the numerical variable) for the *ith* level of the categorical variable.
	When is this analysis *(for the population)* appropriate to use?	**Conditions for ANOVA Inference** 1. Each of the p sample distributions are random. 2. $n_i < 10\%$ of the population distribution. 3. Each sample distribution is approximately normal. 4. $s_1 \approx s_2 \approx \cdots \approx s_p$ 5. Each of the *p* samples are independent of each other.
Predictive Analytics	**Making Predictions**	N/A
	How to quantify the performance of your prediction(s)?	N/A

4. MODELING THE ASSOCIATION BETWEEN A CATEGORICAL EXPLANATORY VARIABLE (WITH >2 LEVELS) AND A NUMERICAL RESPONSE VARIABLE *IN THE SAMPLE*

Ex: We would first like to model the association between political association and age, first using a multiple linear regression model (with age as the response variable and party as the explanatory variable.) In this analysis, we will again examine our Pew Survey dataset (from 2017) which has randomly sampled n=1465 adults living in the U.S., which has five possible values in the "party" variable: Democrat, Independent, No preference, Other party, and Republican.

If we were to set up a multiple linear regression model for this analysis, how many slopes would the model have?

	party	age
0	Democrat	18.0
1	Independent	18.0
2	No preference (VOL.)	19.0
3	Other party (VOL.)	32.0
4	Republican	18.0
	...	...

Using Python we find that the model yields the following output.

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, 30 Mar 2020	**Prob (F-statistic):**	6.66e-16
Time:	10:27:55	**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

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

With the information in the summary output table, we can find the intercept and slopes for this multiple linear regression model *for the sample data*.

$$\widehat{age} = 50.4991 - 3.6914party[T.Ind] - 7.3527party[T.No\_Pref] - 5.8991party[T.Other] + 6.2775party[T.Rep]$$

By interpreting our intercept and slopes, we can use this multiple linear regression model to say the following things about the sample.

- $\underline{\widehat{\beta_0}\ Interpretation}$: We expect the mean age of **Democrats** in the sample to be **50.4991.**

- $\underline{\widehat{\beta_1}\ Interpretation}$: We expect the mean age of **Independents** in the sample to be 3.6914 years lower than the mean age of Democrats in the sample (ie. **46.808**=50.4991-3.6914).

- $\underline{\widehat{\beta_2}\ Interpretation}$: We expect the mean age of people with **no party preference** in the sample to be 7.3527 years lower than the mean age of Democrats in the sample (ie. **43.1464**=50.4991-7.3527).

- $\underline{\widehat{\beta_3}\ Interpretation}$: We expect the mean age of people from **"other parties"** in the sample to be 5.8991 years lower than the mean age of Democrats in the sample (ie. **44.6**=50.4991-5.8991).

- $\underline{\widehat{\beta_4}\ Interpretation}$: We expect the mean age of **Republicans** in the sample to be 6.2775 years higher than the mean age of Democrats in the sample (ie. **56.7766**=50.4991+6.2775).

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

Let's find the actual sample mean age of each of the five political affiliations in this Pew dataset and compare them to what we said when we interpreted our slopes and intercept.

Sample Mean Ages

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

Each of these sample mean ages matches up to what we said when we interpreted the intercepts and slopes of the multiple linear regression line!

In general, this relationship that we observed between the sample slopes and intercepts and the sample means will hold whenever you have a multiple linear regression model with a single explanatory variable that is categorical.

5. MODELING THE ASSOCIATION BETWEEN A CATEGORICAL EXPLANATORY VARIABLE (WITH >2 LEVELS) AND A NUMERICAL RESPONSE VARIABLE *IN THE POPULATION*

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.

$$\widehat{age} = \beta_0 + \beta_1 party[T.Ind] + \beta_2 party[T.Ind] + \beta_3 party[T.No\ pref] + \beta_4 party[T.Rep]$$

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

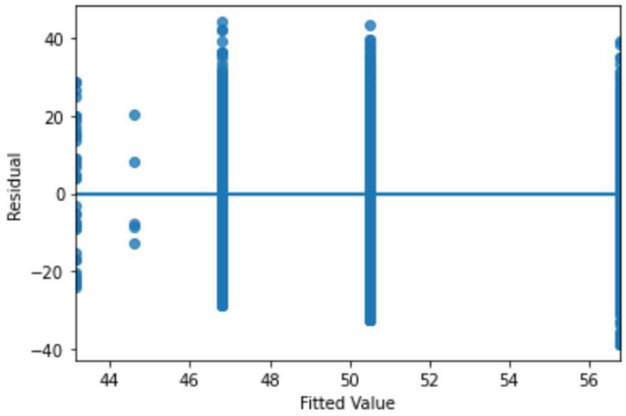
1. **First let's formulate the hypotheses.**

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

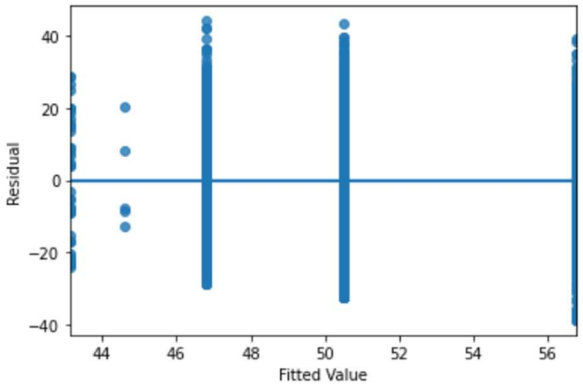
   $H_A: At\ least\ one\ \beta_i \neq 0$ (for i=1,2,3,4)

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

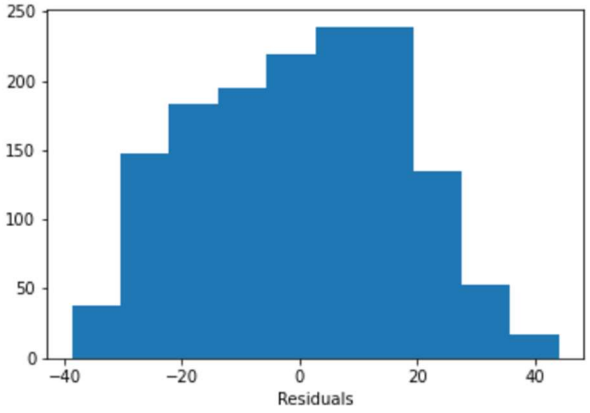
   a. Linearity Condition?



   b. Constant Variability of Residuals Condition?



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



d. Independence of residuals condition?

e. No multicollinearity condition?

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

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:	16:03:02	**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.Independent]	-3.6914	1.073	-3.440	0.001	-5.796	-1.587
party[T.No preference (VOL.)]	-7.3527	2.821	-2.606	0.009	-12.887	-1.818
party[T.Other party (VOL.)]	-5.8991	7.819	-0.754	0.451	-21.237	9.439
party[T.Republican]	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

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

5.1. ANOVA: Relationship between the Population slopes of this multiple linear regression and the Population means

In section 4, we saw that there was a relationship between the **sample** intercept and the 4 sample slopes of the multiple linear regression model ($ie. \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4$) and the 5 **sample** means ages corresponding to the 5 political affiliations (ie. $\bar{x}_{dem}, \bar{x}_{ind}, \bar{x}_{no\ pref}, \bar{x}_{no\ party}, \bar{x}_{rep}$).

Is there also a relationship between the **population** intercept and the 4 sample slopes of the multiple linear regression model ($\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$) and the 5 **population** means ages corresponding to the 5 political affiliations (ie. $\mu_{dem}, \mu_{ind}, \mu_{no\ pref}, \mu_{no\ party}, \mu_{rep}$)?

Yes!

Response Variable: Numerical
Explanatory Variable: Categorical (w levels)

$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{w-1} x_{w-1}$ $\quad\quad$ $\mu_1, \mu_2, \ldots, \mu_w$

Hypotheses

$H_0: \beta_1 = \beta_2 = \cdots = \beta_{w-1} = 0$

H_A: at least one $\beta_i \neq 0$ for i=1,2,...w-1

Test Statistic

$$test\ statistic = \frac{SSR/(w-1)}{SSE/(n-(w-1)-1)}$$

P-value

$$p - value = P(F_{w-1,n-(w-1)-1} \geq test\ statistic)$$

Hypotheses

$H_0: \mu_1 = \mu_2 = \cdots = \mu_w$

$H_A: \mu_i \neq \mu_j$ for at least one pair (i, j) between 1 and w

Test Statistic

P-value

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?

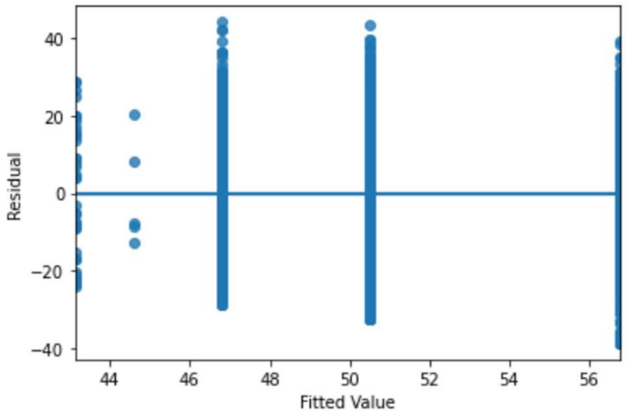
1. **First let's formulate the hypotheses.**

   $H_0: \mu_{dem} = \mu_{rep} = \mu_{ind} = \mu_{no\ pref} = \mu_{no\ party}$

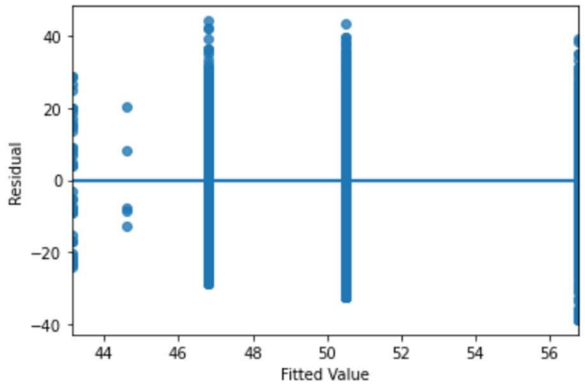
   $H_A: \mu_{group} \neq \mu_{another\ group}\ for\ at\ least\ one\ pair\ of\ political\ affiliations\ groups$

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 $\widehat{age} = \beta_0 + \beta_1 party[T.Ind] + \beta_2 party[T.Ind] + \beta_3 party[T.No\ pref] + \beta_4 party[T.Rep]$)

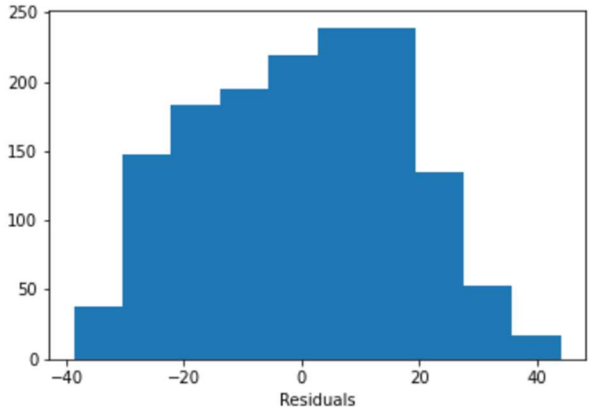
   a. <u>Linearity Condition?</u>



   b. <u>Constant Variability of Residuals Condition?</u>



   c. <u>Normality of Residuals Condition (with Mean 0)?</u>



d. Independence of residuals condition?

e. No multicollinearity condition?

3. **<**

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:	16:03:02	**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.Independent]	-3.6914	1.073	-3.440	0.001	-5.796	-1.587
party[T.No preference (VOL.)]	-7.3527	2.821	-2.606	0.009	-12.887	-1.818
party[T.Other party (VOL.)]	-5.8991	7.819	-0.754	0.451	-21.237	9.439
party[T.Republican]	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

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