POLYNOMIAL REGRESSION WITH A CATEGORICAL PREDICTOR

CATEGORICAL PREDICTORS

Two types of categorical predictors:

- **Nominal**: the predictor variable is composed of categorical options with no inherent ordering
  - Example: Using **state of residence** (Washington, Oregon, California) to predict **housing prices**
  - With this type of predictor, we've focused on testing **differences between group means**

- **Ordinal**: the predictor variable is composed of categorical options with an inherent ordering
  - Example: Using **grade in school** (4^{th}, 5^{th}, 6^{th}) to predict **emotional intelligence**
  - With this type of predictor, we can test **differences between group means**, but we can also test whether there are certain **trends (aka, patterns) in the group means**

TRENDS IN THE GROUP MEANS

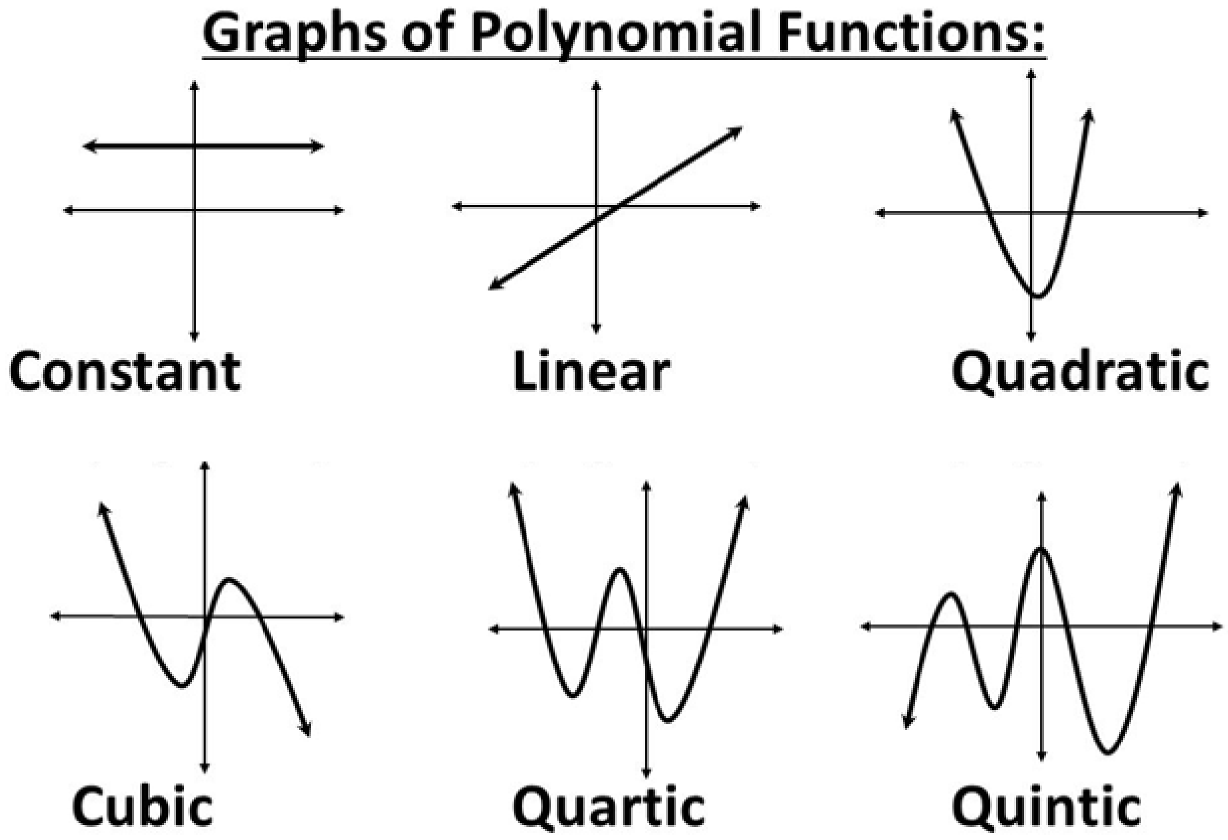
- Example: Say we collect data on which **age group** people belong to as well as how strongly they support **forgiveness of student loans**.

- Age group:
  - Under 18
  - 19-35
  - 36 and older

- There is an inherent ordering to the levels of **age group** (youngest to oldest).

- We could be interested in testing **the pattern of how the means for each condition change**.
  - Specifically, let's say we're interested in whether the pattern of change in the groups' means follows the pattern of a <u>polynomial function</u>.

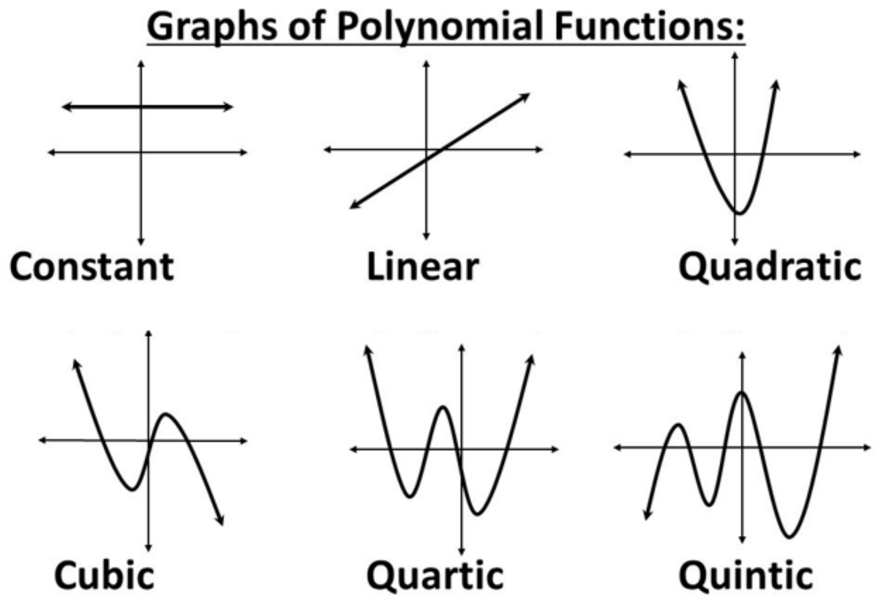
TYPES OF POLYNOMIAL FUNCTIONS



TYPES OF POLYNOMIAL FUNCTIONS

- For each of these polynomial functions, you need the following number of levels of the categorical predictor to test for its trend in the pattern of the group means:

Type of Function	**Number of Levels of Categorical Predictor Needed**
Linear	At least two groups
Quadratic	Three groups
Cubic	Four groups
Quartic	Five groups
Qunitic	Six groups



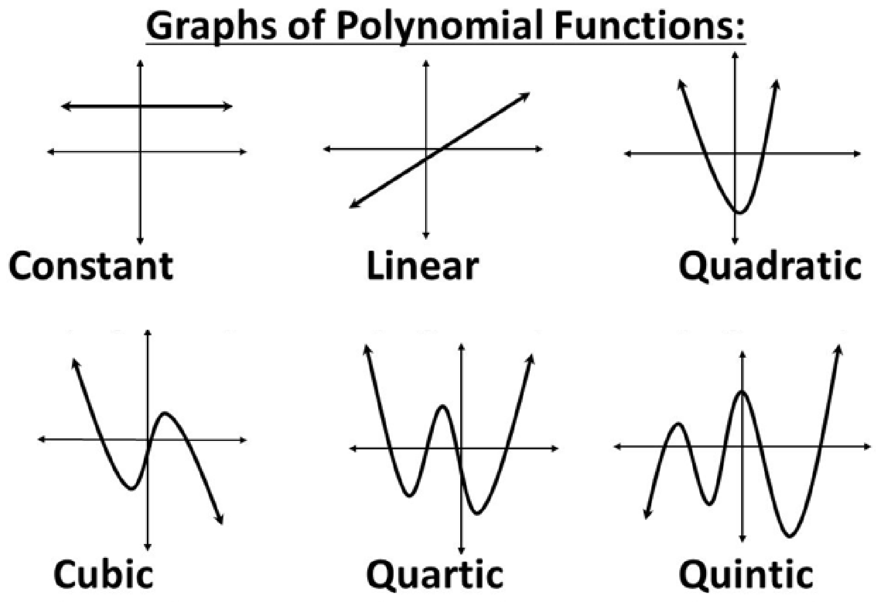
TYPES OF POLYNOMIAL FUNCTIONS

- For our example, we are looking at how average support for the forgiveness of student loans changes across three age groups:

- **Age group:**
  - Under 18
  - 19-35
  - 36 and older

- The categorical predictor has **three groups** so we can test for a linear and a quadratic trend in the pattern of the groups' means

- *Q:* For this scenario, what pattern might you expect the pattern of the change in the groups' means to follow?



TRENDS IN THE GROUP MEANS

- Let's say we collected data from 18 participants, six of whom belonged to each of our three age groups. We asked them, on a scale from 1 to 10, how strongly they support forgiveness of student loans. Their raw scores are shown in the table below.

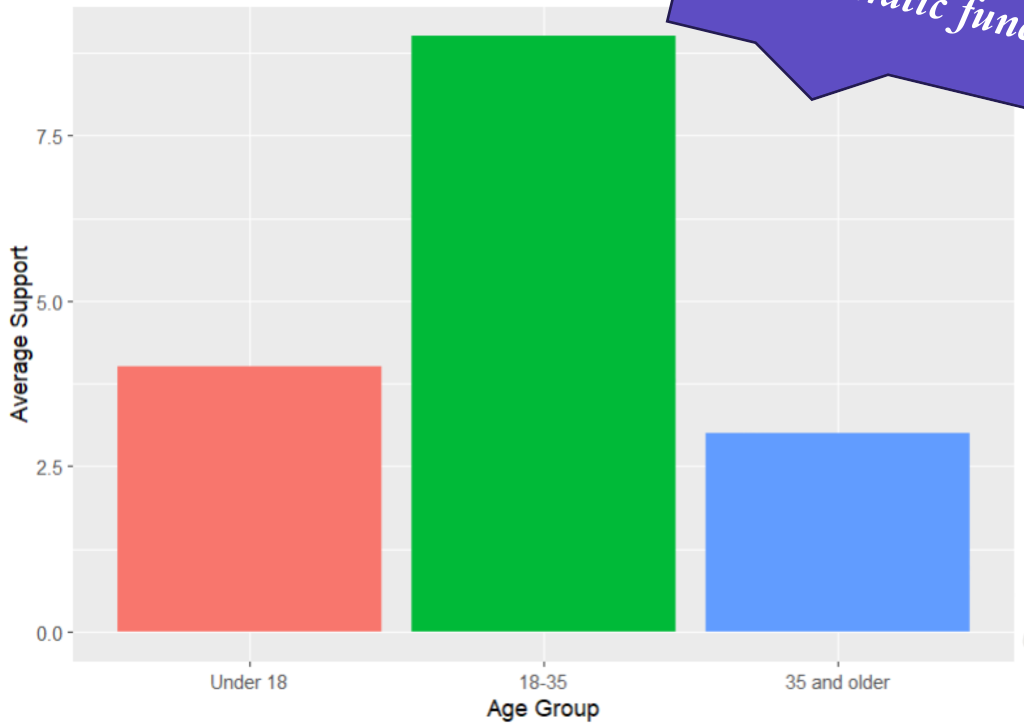
Age Group		
Under 18	19-35	36 and older
4	9	5
3	10	2
5	8	1
6	9	6
3	10	3
3	8	1
$M_1 = 4$	$M_2 = 9$	$M_3 = 3$

- The best way to initially examine whether the pattern of the group means follows a particular polynomial trend is by **graphing the group means**.

TRENDS IN THE GROUP MEANS

- Let's say we collected data from 18 participants, six of whom belonged to each of our three age We asked them, on a scale from 1 to 10, how strongly they support forgiveness of student loans raw scores are shown in the table below.

Age Group		
Under 18	19-35	36 and older
4	9	5
3	10	2
5	8	1
6	9	6
3	10	3
3	8	1
$M_1 = 4$	$M_2 = 9$	$M_3 = 3$



The pattern of change in the groups' means appears to follow the pattern of a ***quadratic function***.

CONTRAST CODING

- For a categorical predictor with **three levels**, we need **two contrast codes**. We can also test for the significance of a **linear trend** and a **quadratic trend** in the data using the following set of contrast codes:

	Under 18	19-35	36 and older
LinearCC	-1	0	1
QuadraticCC	-1	2	-1

Question:

- How does the first contrast code test for a significant linear trend in the pattern of the group means?
- How does the second contrast code test for a significant quadratic trend in the pattern of the group means?

POLYNOMIAL CONTRAST CODES

FIGURE 8.16 Orthogonal polynomial contrast codes

Trend	*Category*				
	1	2			
Linear	–1	1			
	1	2	3		
Linear	–1	0	1		
Quadratic	–1	2	–1		
	1	2	3	4	
Linear	–3	–1	1	3	
Quadratic	1	–1	–1	1	
Cubic	–1	3	–3	1	
	1	2	3	4	5
Linear	–2	–1	0	1	2
Quadratic	2	–1	–2	–1	2
Cubic	–1	2	0	–2	1
Quartic	1	–4	6	–4	1

\# of levels of categorical predictor

contrast codes

MODEL COMPARISON

First, let's test for the significance of a **linear trend** in the groups' means.

Model Comparison

Model A: $Y_i = \beta_0 + \beta_1 \text{LinearCC}_i + \beta_2 \text{QuadraticCC}_i + \varepsilon_i$ PA = 3

Model C: $Y_i = \beta_0 + \beta_2 \text{QuadraticCC}_i + \varepsilon_i$ PC = 2

Null & Alternative Hypotheses:

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

Estimate of Model A and Model C:

Model A: $Y_i = \beta_0 + \beta_1 LinearCC_i + \beta_2 QuadraticCC_i + \varepsilon_i$

Estimate of Model A:

- $Y_i = 5.33 + -0.5 * LinearCC_i + 1.83 * QuadraticCC_i$

Model C: $Y_i = \beta_0 + \beta_2 QuadraticCC_i + \varepsilon_i$

Estimate of Model C:

- $Y_i = 5.33 + 1.83 * QuadraticCC_i$

- Which model fits the data better?
  - Compare fit of Model A to fit of Model C

Age Group		
Under 18	19-35	36 and older
4	9	5
3	10	2
5	8	1
6	9	6
3	10	3
3	8	1
$M_1 = 4$	$M_2 = 9$	$M_3 = 3$

	Under 18	19-35	36 and older
LinearCC	-1	0	1
QuadraticCC	-1	2	-1

$$SSE(A) = \Sigma(Y_i - Y'_i)^2$$

	Under 18	19-35	36 and older
LinearCC	-1	0	1
QuadraticCC	-1	2	-1

Y_i	$Y' = 5.33 + -0.5*LinearCC_i + 1.83*QuadraticCC_i$	$(Y_i - Y')$	$(Y_i - Y')^2$
4	5.33 + (-0.5*-1) + (1.83*-1) = 4	4 – 4 = 0	0
3	4	-1	1
5	4	1	1
6	4	2	4
3	4	-1	1
3	4	-1	1
9	5.33 + (-0.5*0) + (1.83*2) = 9	9 – 9 = 0	0
10	9	1	1
8	9	-1	1
9	9	0	0
10	9	1	1
8	9	-1	1
5	5.33 + (-0.5*1) + (1.83*-1) = 3	5 – 3 = 2	4
2	3	-1	1
1	3	-2	4
6	3	3	9
3	3	0	0
1	3	-2	4

SSE(A) = 34

$$SSE(C) = \Sigma(Y_i - Y'_i)^2$$

	Under 18	19-35	36 and older
LinearCC	-1	0	1
QuadraticCC	-1	2	-1

Y_i	$Y' = 5.33 + 1.83*QuadraticCC_i$	$(Y_i - Y')$	$(Y_i - Y')^2$
4	5.33 + (1.83*-1) = 3.5	4 – 3.5 = 0.5	0.25
3	3.5	-0.5	0.25
5	3.5	1.5	2.25
6	3.5	2.5	6.25
3	3.5	-0.5	0.25
3	3.5	-0.5	0.25
9	5.33 + (1.83*2) = 9	0	0
10	9	1	1
8	9	-1	1
9	9	0	0
10	9	1	1
8	9	-1	1
5	5.33 + (1.83*-1) = 3.5	1.5	2.25
2	3.5	-1.5	2.25
1	3.5	-2.5	6.25
6	3.5	2.5	6.25
3	3.5	-0.5	0.25
1	3.5	-2.5	6.25

SSE(C) = 37

CALCULATING F-STATISTIC

	SS	***df***	***MS***	***F***	***p***
LinearCC	SSR = 37-34 = 3	PA-PC = 3-2 = 1	MSR = 3/1 = 3	F = 3/2.2667 = 1.32	Use R to obtain
QuadraticCC					
Model A	SSE(A) = 34	n – PA = 18-3 = 15	MSE = 34/15 = 2.2667		

MODEL COMPARISON

Second, let's test the significance of a **quadratic trend** in the groups' means.

<u>Model Comparison</u>

Model A: $Y_i = \beta_0 + \beta_1 \text{LinearCC}_i + \beta_2 \text{QuadraticCC}_i + \varepsilon_i$ PA = 3

Model C: $Y_i = \beta_0 + \beta_1 \text{LinearCC}_i + \varepsilon_i$ PC = 2

<u>Null & Alternative Hypotheses:</u>

H_0: $\beta_2 = 0$

H_1: $\beta_2 \neq 0$

Estimate of Model A and Model C:

Model A: $Y_i = \beta_0 + \beta_1 \text{LinearCC}_i + \beta_2 \text{QuadraticCC}_i + \varepsilon_i$

Estimate of Model A:

- $Y_i = 5.33 + -0.5 \star \text{LinearCC}_i + 1.83 \star \text{QuadraticCC}_i$
  - Model A hasn't changed!

Model C: $Y_i = \beta_0 + \beta_1 \text{LinearCC}_i + \varepsilon_i$

Estimate of Model C:

- $Y_i = 5.33 + -0.5 \star \text{LinearCC}_i$

Age Group		
Under 18	19-35	36 and older
4	9	5
3	10	2
5	8	1
6	9	6
3	10	3
3	8	1
$M_1 = 4$	$M_2 = 9$	$M_3 = 3$

	Under 18	19-35	36 and older
LinearCC	-1	0	1
QuadraticCC	-1	2	-1

$$SSE(C) = \Sigma(Y_i - Y'_i)^2$$

	Under 18	19-35	36 and older
LinearCC	-1	0	1
QuadraticCC	-1	2	-1

Y_i	$Y' = 5.33 + -0.5*LinearCC_i$	$(Y_i - Y')$	$(Y_i - Y')^2$
4	5.33 + (-0.5*-1) = 5.83	-1.83	3.3489
3	5.83	-2.83	8.0089
5	5.83	-0.83	0.6889
6	5.83	0.17	0.0289
3	5.83	-2.83	8.0089
3	5.83	-2.83	8.0089
9	5.33 + (-0.5*0) = 5.33	3.67	13.4689
10	5.33	4.67	21.8089
8	5.33	2.67	7.1289
9	5.33	3.67	13.4689
10	5.33	4.67	21.8089
8	5.33	2.67	7.1289
5	5.33 + (-0.5*1) = 4.83	0.17	0.0289
2	4.83	-2.83	8.0089
1	4.83	-3.83	14.6689
6	4.83	1.17	1.3689
3	4.83	-1.83	3.3489
1	4.83	-3.83	14.6689

$SSE(C) = 155$

CALCULATING F-STATISTIC

	SS	***df***	***MS***	***F***	***p***
LinearCC	SSR = 37-34 = 3	PA-PC = 3-2 = 1	MSR = 3/1 = 3	F = 3/2.2667 = 1.32	Use R to obtain
QuadraticCC	SSR = 155-34 = 121	PA-PC = 3-2 = 1	MSR = 121/1 = 121	F = 121/2.2667 = 53.38	Use R to obtain
Model A	SSE(A) = 34	n – PA = 18-3 = 15	MSE = 34/15 = 2.2667		

PERFORMING THE ANALYSIS IN R

- anova( ) table

```
Analysis of Variance Table

Response: forgiveness
          Df Sum Sq Mean Sq F value    Pr(>F)    
linearCC   1      3   3.000  1.3235     0.268    
quadCC     1    121 121.000 53.3824 2.583e-06 ***
Residuals 15     34   2.267                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- There was not a significant linear trend in the pattern of the group means, $F(1, 15) = 1.32$, $p = .268$.
- There was a significant quadratic trend in the pattern of the group means, $F(1, 15) = 53.38$, $p < .001$.

PERFORMING THE ANALYSIS IN R

- summary( ) table

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   5.3333     0.3549  15.029 1.89e-10 ***
linearCC     -0.5000     0.4346  -1.150    0.268    
quadCC        1.8333     0.2509   7.306 2.58e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.506 on 15 degrees of freedom
Multiple R-squared:  0.7848,	Adjusted R-squared:  0.7561 
F-statistic: 27.35 on 2 and 15 DF,  p-value: 9.912e-06
```

- The full model including the linear and quadratic trends for age groups accounted for a significant amount of variance in people's support for the forgiveness of student loans, $R^2 = 0.78$, $F(2, 15) = 27.35$, $p < .001$.

- Specifically, there was no significant linear trend, $b = -0.50$, $t(15) = -1.15$, $p = .268$, but there was a significant quadratic trend in the pattern of the group means, $b = 1.83$, $t(15) = 7.31$, $p < .001$.