Advanced Regression: Interaction Terms, Polynomial Terms, and Fixed Effects

Rafael Campos-Gottardo[*]

2025-03-26

Learning Objectives

- Understand interactions terms and how to interpret marginal effects.
- Visualize polynomial terms in regression.
- Understand fixed effects as isolating within variation by removing between variation.

Before Lab

Before your lab this week please ensure that you have read the lab guide and downloaded the data into your project folder you will also be more prepared for the lab if you complete the following:

- Read pages 19-22 of the Franzese and Kam reading (on *mycourses*).
- Read section 13.2.2 Polynomials of *The Effect*.
- Read Chapter 16 of the *The Effect*.

Getting started

Before we start we need to install the marginal effects package. This package is one of the most important R packages for summarizing and interpreting complex statistical models.[1]

```r
#install.packages("marginaleffects")
library(marginaleffects)
```

This week we are going to use our class version of the Canadian Election Study (CES) again. Let's load the data set and then pick a model we wish to estimate (remember the dataset is in the `Stata` format so we need to use the `haven` package to load it).

```r
library(haven)
CES <- read_dta("poli311_CES.dta")
head(CES)
```

```
## # A tibble: 6 x 6
##     Age Province         DemSat Gender  VoteChoice       Ideology
##   <dbl> <chr>             <dbl> <chr>   <chr>               <dbl>
## 1    57 Quebec                3 A Man   ""                      6
## 2    22 British Columbia      3 A Woman "NDP"                   3
## 3    28 British Columbia      3 A Woman ""                      5
## 4    29 Ontario               4 A Woman ""                      0
## 5    41 Quebec                3 A Woman "NDP"                   4
## 6    63 Quebec                3 A Woman "Bloc Quebecois"        0
```

[*] This lab guide builds on those created by Aaron Erlich and Colin Scott. I would also like to acknowledge Brendan Szendro, Nicholas Vaxelaire, Guila Cohen, and Natalie Delmonte for their help developing this lab guide.

[1] As an added bonus it was developed by a local professor at the University of Montreal - Vincent Arel-Bundock.

For ease of demonstration we are first going to filter out the non-binary respondents. We have done this a number of times in lab so try to do this code on your own before looking at the answer.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
CES <- CES %>%
 filter(Gender %in% c("A Man", "A Woman"))
```

Let us now estimate the relationship between age and ideology while controlling for Gender and satisfaction with democracy. We can formally write this model in the following way:[2]

$$Y_i = \beta_0 + \beta_1 Age_i + \beta_2 Gender_i + \beta_3 DemSat + \varepsilon_i$$

Linear regression model

Now that we have prepared our data set we can estimate our basic linear regression model.

```
model_1 <- lm(Ideology ~ Age + Gender + DemSat, data = CES)
```

We can summarize our model using the `modelsummary` package discussed in last week's lab guide.

```
#install.packages("modelsummary")
library(modelsummary)
```

```
## `modelsummary` 2.0.0 now uses `tinytable` as its default table-drawing
##   backend. Learn more at: https://vincentarelbundock.github.io/tinytable/
##
## Revert to `kableExtra` for one session:
##
##   options(modelsummary_factory_default = 'kableExtra')
##   options(modelsummary_factory_latex = 'kableExtra')
##   options(modelsummary_factory_html = 'kableExtra')
##
## Silence this message forever:
##
##   config_modelsummary(startup_message = FALSE)
```

```
modelsummary(model_1, stars = TRUE)
```

We can see that in this model age is significantly related to satisfaction with democracy where older individuals are right-wing. Specifically, a one-year increase in age is associated with a 0.014 unit increase in being more right-wing. Additionally, women are less right-wing than men and those who are more satisfied with democracy are less right-wing. However, we may have reason to believe that the relationship between age and ideology varies by gender. In other words the relationship between age and ideology is conditional on a respondent's gender.

[2] For you assignments we expect that you write out your main model in this format.

	(1)
(Intercept)	5.365***
	(0.087)
Age	0.014***
	(0.001)
GenderA Woman	−0.360***
	(0.035)
DemSat	−0.287***
	(0.023)
Num.Obs.	17 697
R2	0.028
R2 Adj.	0.028
AIC	79 874.1
BIC	79 913.0
Log.Lik.	−39 932.052
RMSE	2.31

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Interaction Terms

One way we can test whether the relationship between age and ideology varies based on gender is by doing sub-group analysis.

We can fit a separate regression model for men and women.

```
# Men
model_2_men <- lm(Ideology ~ Age + DemSat, data = CES %>% filter(Gender == "A Man"))
# Notice how we can use the pipe within a function to make filtering easier
model_2_women <- lm(Ideology ~ Age + DemSat, data = CES %>% filter(Gender == "A Woman"))
```

Now that we have fitted these two models we can use the `modelsummary()` function to display them next to each other along with our original model.

```
modelsummary(list("Pooled Model" = model_1,
                  # You can name your models to clean up your output
                  "Men Only" = model_2_men,
                  "Women Only" = model_2_women),
             stars = TRUE)
```

In this model we can see that age has a higher correlation with ideology for women then men with the age coefficient being larger for women (0.019 > 0.010). However, by doing a sub-group analysis we do not know whether this difference in coefficients (effect size if this was a causal study) is statistically significant. In order to determine if this difference is statistically significant we can fit a new model using an interaction term.

This model can be formally written as follows:

$$Y_i = \beta_0 + \beta_1 Age_i + \beta_2 Gender_i + \beta_3 (Age_i \times Gender_i) + \beta_4 DemSat + \varepsilon_i$$

This model allows the slope (coefficient) of both Age and Gender to vary depending on the value of the other.

	Pooled Model	Men Only	Women Only
(Intercept)	5.365***	5.638***	4.801***
	(0.087)	(0.117)	(0.121)
Age	0.014***	0.010***	0.019***
	(0.001)	(0.001)	(0.001)
GenderA Woman	−0.360***		
	(0.035)		
DemSat	−0.287***	−0.288***	−0.287***
	(0.023)	(0.030)	(0.035)
Num.Obs.	17 697	8681	9016
R2	0.028	0.014	0.023
R2 Adj.	0.028	0.014	0.023
AIC	79 874.1	38 548.6	41 271.0
BIC	79 913.0	38 576.9	41 299.4
Log.Lik.	−39 932.052	−19 270.311	−20 631.485
RMSE	2.31	2.23	2.39

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

In order to, implement an interaction using R we add a multiplication symbol between the variables we want to interact in our model. [Note that you could also use the colon symbol (:) to implement an interaction, however, R will only provide a coefficient for the interaction and not for the individual variables.]

```
interaction_model <- lm(Ideology ~ Age*Gender + DemSat, data = CES)
```

Now we can add this model to our model summary.

```
modelsummary(list("Pooled Model" = model_1,
                  "Men Only" = model_2_men,
                  "Women Only" = model_2_women,
                  "Interaction Model" = interaction_model),
             stars = TRUE)
```

What do you notice about the output from this model?

1. Notice that we get a new coefficient in this model `Age x GenderA Woman` this coefficient is called an interaction term and tells us whether the difference in our coefficients is statistically significant. Since our interaction term has three stars we can say that there is statistically significant difference in the effect of age on ideology between men and women.
2. Notice that our age coefficient is the same as the age coefficient for our men only model. This is because the age coefficient represents the relationship between age and ideology for the baseline group for our interaction term (in this case men). In order to get the coefficient for women we need to add the interaction term to the age coefficient $0.010 + 0.009 = 0.019$. Adding these coefficients provides the coefficient for the relationship between age and ideology for women.

Another way we can present these results is by using `slopes()` function from the marginal effects package. This function provides the coefficients for age for each level of gender.

```
slopes(interaction_model, variables = "Age", by = "Gender")
```

```
##
##   Term     Contrast   Gender Estimate Std. Error      z Pr(>|z|)     S   2.5 %
```

	Pooled Model	Men Only	Women Only	Interaction Model
(Intercept)	5.365***	5.638***	4.801***	5.637***
	(0.087)	(0.117)	(0.121)	(0.107)
Age	0.014***	0.010***	0.019***	0.010***
	(0.001)	(0.001)	(0.001)	(0.002)
GenderA Woman	−0.360***			−0.834***
	(0.035)			(0.114)
DemSat	−0.287***	−0.288***	−0.287***	−0.287***
	(0.023)	(0.030)	(0.035)	(0.023)
Age × GenderA Woman				0.009***
				(0.002)
Num.Obs.	17 697	8681	9016	17 697
R2	0.028	0.014	0.023	0.029
R2 Adj.	0.028	0.014	0.023	0.029
AIC	79 874.1	38 548.6	41 271.0	79 857.0
BIC	79 913.0	38 576.9	41 299.4	79 903.7
Log.Lik.	−39 932.052	−19 270.311	−20 631.485	−39 922.510
RMSE	2.31	2.23	2.39	2.31

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

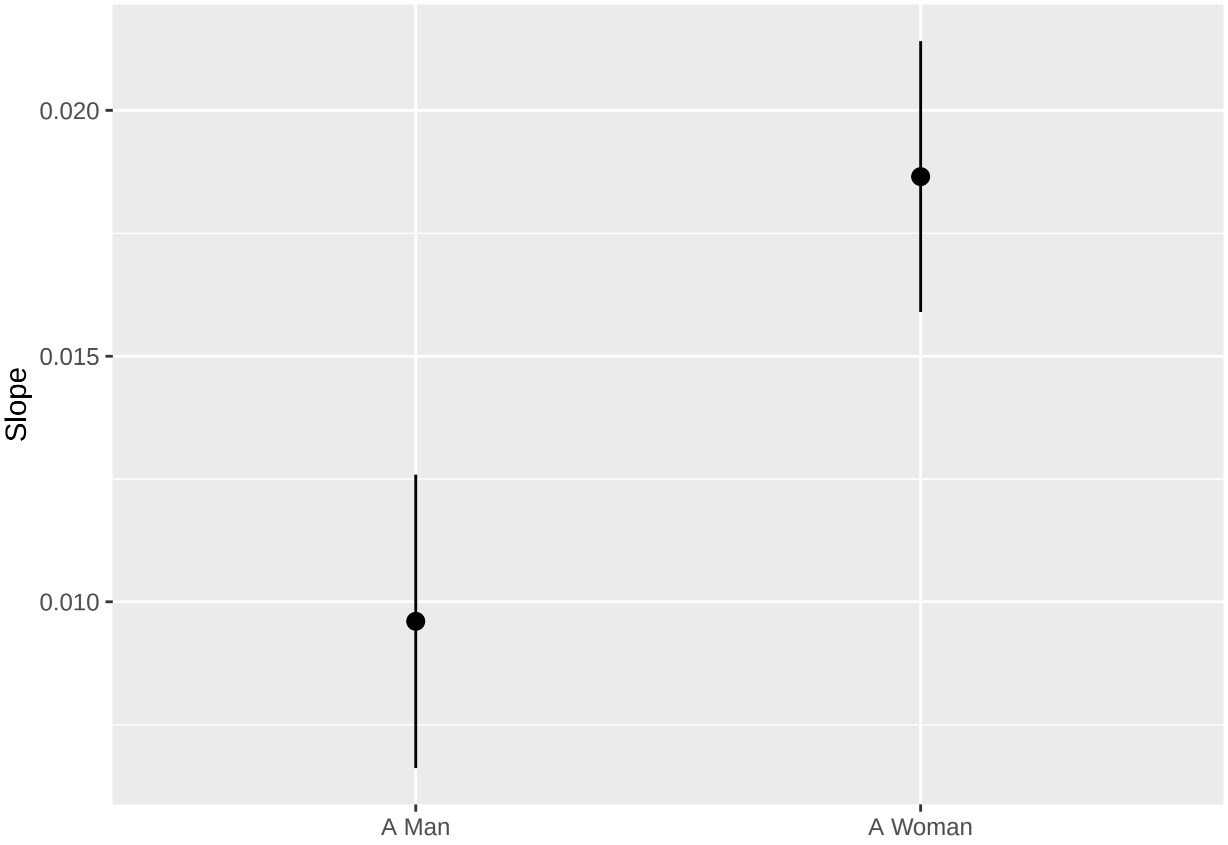
```
##   Age mean(dY/dX) A Man    0.00961    0.00152  6.31   <0.001  31.7 0.00662
##   Age mean(dY/dX) A Woman  0.01865    0.00141 13.26   <0.001 130.9 0.01590
##  97.5 %
##  0.0126
##  0.0214
##
## Columns: term, contrast, Gender, estimate, std.error, statistic, p.value, s.value, conf.low, conf.hi
## Type:  response
```

Notice how we get the same coefficients as in our sub-group models and our interaction model.

Thought Question: What happens if we put "Gender" in the variables argument and "Age" in the by argument?

We can also use the `plot_slopes()` function from the marginal effects package to plot these regression coefficients and the difference. FYI this function produces a "gg" object that we can add additional `ggplot` arguments to.

```r
plot_slopes(interaction_model, variables = "Age", by = "Gender")
```



Polynomial terms

Another transformation we can include in our regression models is polynomial terms. In a standard linear regression we assume that the relationship between our X and Y variables is linear. However, for many social phenomena the relationship between X and Y is not linear. Therefore, we can include a polynomial term to let our regression line curve. One variable that is often not linearly related to dependent variables is age.

We can fit a new model that examines the relationship between age and ideology with a polynomial term that can be formally expressed as so:

$$Y_i = \beta_0 + \beta_1 Age_i + \beta_2 Age_i^2 + \beta_3 Gender_i + \beta_4 DemSat + \varepsilon_i$$

In order to fit this model we will use the `I()` function within our regression model which allows us to transform variables within the model using mathematical operators. It is important to also keep the original age coefficient in our model to retain the original linear relationship.

```
polynomial_model <- lm(Ideology ~ Age + I(Age^2) + Gender + DemSat, data = CES)
```

Let's also add this model to our model summary.

```
modelsummary(list("Pooled Model" = model_1,
                  "Men Only" = model_2_men,
                  "Women Only" = model_2_women,
                  "Interaction Model" = interaction_model,
                  "Curvilinear Model" = polynomial_model),
             stars = TRUE)
```

This model indicates that as age increases individuals get become more right-wing, and that this relationship becomes stronger as people get older. One way to better understand this model is to graph the predictions

	Pooled Model	Men Only	Women Only	Interaction Model	Curvilinear Model
(Intercept)	5.365***	5.638***	4.801***	5.637***	4.650***
	(0.087)	(0.117)	(0.121)	(0.107)	(0.167)
Age	0.014***	0.010***	0.019***	0.010***	0.046***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.006)
GenderA Woman	−0.360***			−0.834***	−0.356***
	(0.035)			(0.114)	(0.035)
DemSat	−0.287***	−0.288***	−0.287***	−0.287***	−0.282***
	(0.023)	(0.030)	(0.035)	(0.023)	(0.023)
Age × GenderA Woman				0.009***	
				(0.002)	
I(Age^2)					0.000***
					(0.000)
Num.Obs.	17 697	8681	9016	17 697	17 697
R2	0.028	0.014	0.023	0.029	0.029
R2 Adj.	0.028	0.014	0.023	0.029	0.029
AIC	79 874.1	38 548.6	41 271.0	79 857.0	79 850.9
BIC	79 913.0	38 576.9	41 299.4	79 903.7	79 897.6
Log.Lik.	−39 932.052	−19 270.311	−20 631.485	−39 922.510	−39 919.439
RMSE	2.31	2.23	2.39	2.31	2.31

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

from it. First we need to create predictions using the predict function.

For ease we are going to re-run the model without controls. However, if you want to include controls when predicting the common practice is to hold the control variables at their mean or median.

```r
polynomial_model_2 <- lm(Ideology ~ Age + I(Age^2), data = CES)
summary(polynomial_model_2)
```

```
##
## Call:
## lm(formula = Ideology ~ Age + I(Age^2), data = CES)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3757 -1.5050 -0.1649  1.6618  5.6422
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.5011255  0.1487996  23.529  < 2e-16 ***
## Age          0.0519154  0.0062969   8.245  < 2e-16 ***
## I(Age^2)    -0.0003594  0.0000617  -5.825 5.81e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.325 on 17962 degrees of freedom
##   (2879 observations deleted due to missingness)
## Multiple R-squared:  0.01513,    Adjusted R-squared:  0.01502
## F-statistic: 137.9 on 2 and 17962 DF,  p-value: < 2.2e-16
```

Notice that when not controlling for Gender and satisfaction with democracy the Age squared coefficient becomes negative. Therefore, our graph is going to have an inverted U-shape.

To create predictions we use the `predict()` function.

```r
polynomial_predictions <- predict(polynomial_model_2,
                                  newdata = data.frame(Age = 18:99))

polynomial_predictions
```

```
##        1        2        3        4        5        6        7        8
## 4.319146 4.357762 4.395660 4.432839 4.469298 4.505039 4.540061 4.574365
##        9       10       11       12       13       14       15       16
## 4.607949 4.640814 4.672961 4.704389 4.735098 4.765088 4.794359 4.822911
##       17       18       19       20       21       22       23       24
## 4.850745 4.877859 4.904255 4.929932 4.954890 4.979129 5.002649 5.025450
##       25       26       27       28       29       30       31       32
## 5.047533 5.068896 5.089541 5.109467 5.128674 5.147162 5.164932 5.181982
##       33       34       35       36       37       38       39       40
## 5.198314 5.213927 5.228820 5.242995 5.256452 5.269189 5.281207 5.292507
##       41       42       43       44       45       46       47       48
## 5.303088 5.312949 5.322092 5.330516 5.338222 5.345208 5.351476 5.357024
##       49       50       51       52       53       54       55       56
## 5.361854 5.365965 5.369357 5.372030 5.373984 5.375220 5.375736 5.375534
##       57       58       59       60       61       62       63       64
## 5.374613 5.372973 5.370614 5.367537 5.363740 5.359224 5.353990 5.348037
##       65       66       67       68       69       70       71       72
## 5.341365 5.333974 5.325864 5.317036 5.307488 5.297222 5.286237 5.274532
##       73       74       75       76       77       78       79       80 
## 5.262109 5.248968 5.235107 5.220527 5.205229 5.189212 5.172476 5.155021 
##       81       82 
## 5.136847 5.117954
```

Now we have predicted values of ideology for individuals aged 18 to 99. Lets create a dataframe so that we can graph this using `ggplot`.

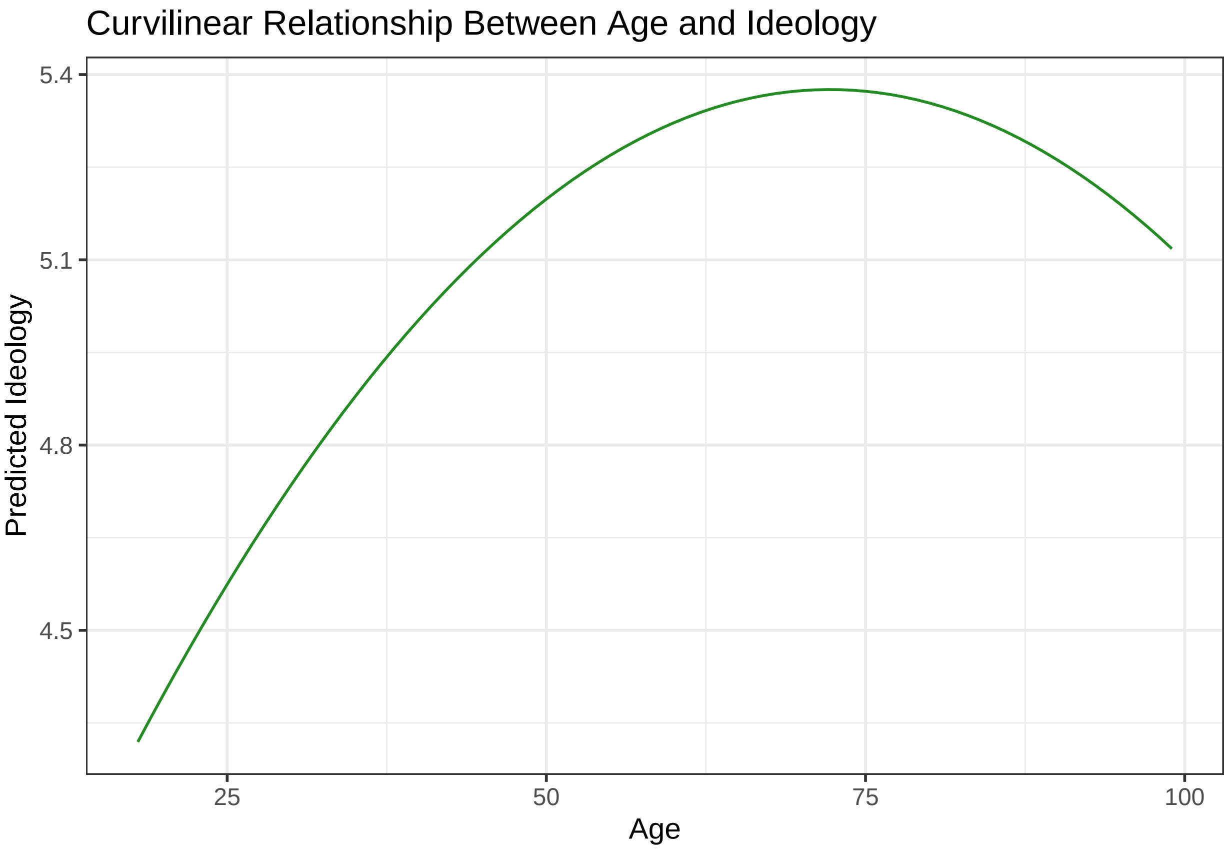
```
predictions_df <- data.frame(Age = 18:99,
                             Predicted_Ideology = polynomial_predictions)

head(predictions_df)
```

```
##   Age Predicted_Ideology
## 1  18           4.319146
## 2  19           4.357762
## 3  20           4.395660
## 4  21           4.432839
## 5  22           4.469298
## 6  23           4.505039
```

Now we can graph this curvilinear relationship.

```
predictions_df %>%
  ggplot(aes(x = Age, y = Predicted_Ideology)) +
  geom_line(colour = "forestgreen") +
  labs(title = "Curvilinear Relationship Between Age and Ideology",
       x = "Age",
       y = "Predicted Ideology") +
  theme_bw()
```



Here you can see that as individuals get older they become more right-wing until they reach their 70s where they start becoming more left-wing again.

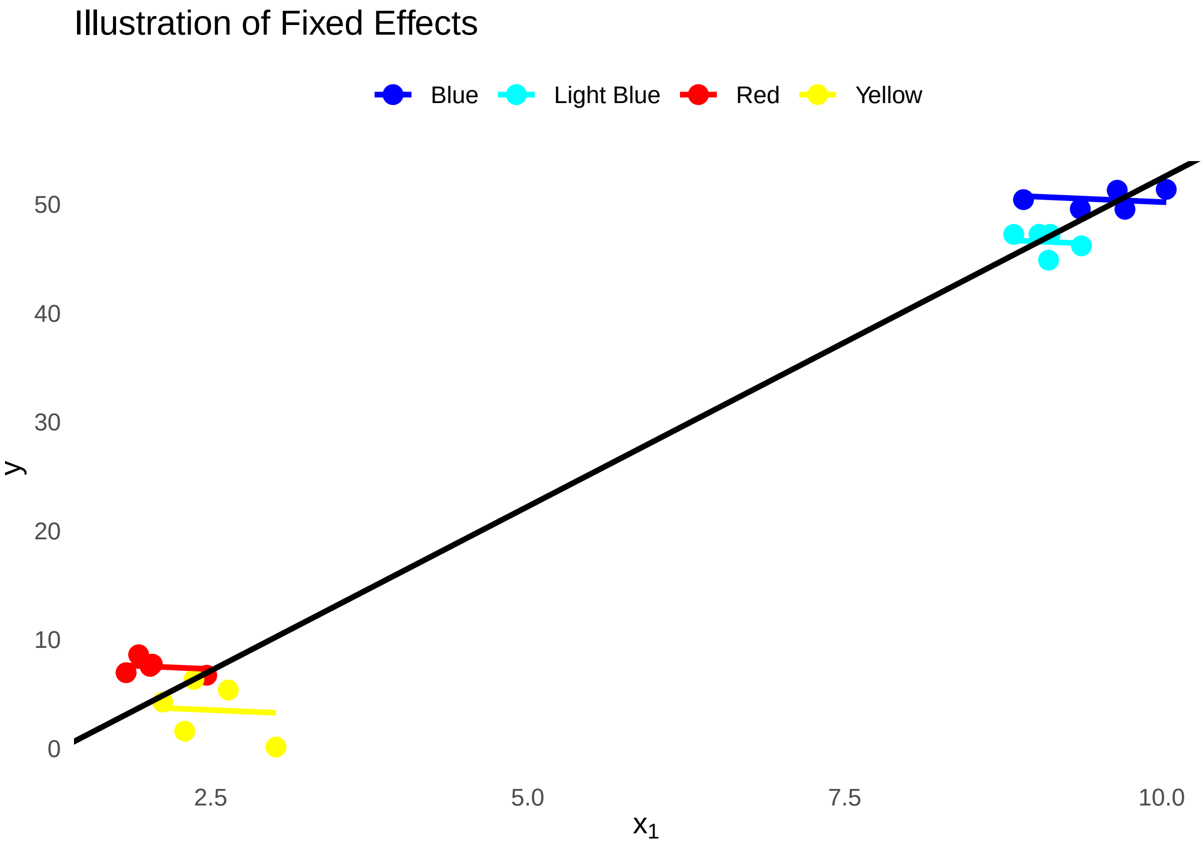
Fixed effects

Before we talk about fixed effects we need to understand between and within variation.

1. Between(-group) variation: The difference across units (Comparing two countries, provinces, individuals, etc.).
2. Within(-group) variation: The differences within a unit across time (changes in levels of polarization within Alberta from 2001-2025). We can also measure within variation without panel data by ignoring the fixed characteristics of a given unit without the time component.

Fixed effects help isolate within-unit variation by removing between-unit variation. We do this by adding unit-level dummy variables (e.g., provinces):

Fixed effects are most common when using data that includes units and time. For example, using fixed effects is useful for country-year data. The inclusion of fixed effects allows a researcher to isolate within variation. For example we might have a panel of Canadian voters and want isolate the relationship between economic globalization and levels of polarization. However, different provinces have different base levels of polarization that do not vary over time. We do not want these time invariant factors to interfere with our analysis so we include fixed effects to remove the between unit variation. Therefore, we can understand fixed effects as helping us remove between variation so that we can isolate the within variation.



This graph demonstrates an extreme example of how our estimate of the relationship between an X and Y variable would be biased due to between variation if we did not include fixed effects in our model. Notice that the "main effect" is positive but the slope for the within variation is negative.

In order to add fixed effects to a regression model we can add dummy variables for each unit. For our model estimating the relationship between age and ideology we can add province fixed effects by including our province variable as a factor variable. This model can be formally written as follows where α_p represents province-level fixed effects (notice they are for p not respondent i):

$$Y_i = \beta_1 Age_i + \beta_2 Age_i^2 + \beta_3 Gender_i + \beta_4 DemSat + \alpha_p + \varepsilon_i$$

We can use R to add fixed effects as follows.

```
fixed_effects_model <- lm(Ideology ~ Age + Gender + DemSat + Province, data = CES)
```

Now we can add our final model to our model summary.

```
modelsummary(list("Pooled Model" = model_1,
                  "Interaction Model" = interaction_model,
                  "Curvilinear Model" = polynomial_model,
                  "Fixed Effects Model" = fixed_effects_model),
             stars = TRUE)
```

The final model only shows us the relationship between age and ideology within each province. Notice that the coefficient is the same as in our baseline model. This indicates that the relationship between age and ideology does not vary across province. However, the coefficients for gender and satisfaction with democracy are different indicating that this relationship does vary across province.

	Pooled Model	Interaction Model	Curvilinear Model	Fixed Effects Model
(Intercept)	5.365***	5.637***	4.650***	5.628***
	(0.087)	(0.107)	(0.167)	(0.097)
Age	0.014***	0.010***	0.046***	0.014***
	(0.001)	(0.002)	(0.006)	(0.001)
GenderA Woman	−0.360***	−0.834***	−0.356***	−0.383***
	(0.035)	(0.114)	(0.035)	(0.035)
DemSat	−0.287***	−0.287***	−0.282***	−0.261***
	(0.023)	(0.023)	(0.023)	(0.023)
Age × GenderA Woman		0.009***		
		(0.002)		
I(Age^2)			0.000***	
			(0.000)	
ProvinceBritish Columbia				−0.362***
				(0.071)
ProvinceManitoba				−0.127
				(0.103)
ProvinceNew Brunswick				−0.581***
				(0.138)
ProvinceNewfoundland and Labrador				−0.214
				(0.194)
ProvinceNorthwest Territories				−1.120+
				(0.667)
ProvinceNova Scotia				−0.636***
				(0.122)
ProvinceNunavut				0.726
				(1.153)
ProvinceOntario				−0.275***
				(0.057)
ProvincePrince Edward Island				0.483
				(0.323)
ProvinceQuebec				−0.480***
				(0.059)
ProvinceSaskatchewan				0.216+
				(0.130)
ProvinceYukon				−0.299
				(0.463)
Num.Obs.	17 697	17 697	17 697	17 697
R2	0.028	0.029	0.029	0.034
R2 Adj.	0.028	0.029	0.029	0.033
AIC	79 874.1	79 857.0	79 850.9	79 784.0
BIC	79 913.0	79 903.7	79 897.6	79 916.3
Log.Lik.	−39 932.052	−39 922.510	−39 919.439	−39 874.996
RMSE	2.31	2.31	2.31	2.30

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Key Takeaways

- Use *interaction terms* to capture how the effect of one variable depends on another.
- Use *polynomial terms* to model non-linear relationships. These are mostly used for curviliear relationships.
- Use *fixed effects* to control unobserved differences across units and focus on within-group variation.