

# Tutorial 14: Working with grouped data

*Johannes Karreth*

*RPOS 517, Day 14*

**This tutorial shows you:**

- how to handle grouped data in R
- how to include fixed effects for groups in regression models

**Note on copying & pasting code from the PDF version of this tutorial:** Please note that you may run into trouble if you copy & paste code from the PDF version of this tutorial into your R script. When the PDF is created, some characters (for instance, quotation marks or indentations) are converted into non-text characters that R won't recognize. To use code from this tutorial, please type it yourself into your R script or you may copy & paste code from the *source file* for this tutorial which is posted on my website.

**Note on R functions discussed in this tutorial:** I don't discuss many functions in detail here and therefore I encourage you to look up the help files for these functions or search the web for them before you use them. This will help you understand the functions better. Each of these functions is well-documented either in its help file (which you can access in R by typing `?ifelse`, for instance) or on the web. The *Companion to Applied Regression* (see our syllabus) also provides many detailed explanations.

As always, please note that this tutorial only accompanies the other materials for Day 14 (in this case, the course video linked on the course website) and that you need to have worked through the materials for that day before tackling this tutorial. **More than on the other days of our seminar so far, these notes only scratch the surface of the issues arising with grouped data.** I strongly encourage you to self-study the theory behind fixed effects before using them in your own work. Two recent articles on the topic are worth a look; they include references to other canonical articles on fixed effects, so take a look - even though these articles go beyond the treatment of FEs in this tutorial:

- Bell, A. and Jones, K. (2015). “[Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data.](#)” *Political Science Research and Methods* 3(1): 133–153.
- Clark, T. S. and Linzer, D. A. (Forthcoming). “[Should I use fixed or random effects?](#)” *Political Science Research and Methods*.

Both of these articles point you to textbook treatments of fixed effects in grouped data; you must consult at least one of these textbooks (e.g., Greene, Wooldridge) before using fixed effects in your work.

## Grouped data and the OLS assumptions

In the [tutorial for Day 7](#), you already encountered the two major types of grouped data in social science:

1. individuals nested in higher-level units
  - Example: survey respondents in an international survey are “nested” in countries (500 respondents in the U.S., 500 in Canada, etc.)
2. observations from the same unit observed over at least 2 time periods (“time-series cross-sectional data”)
  - Example: yearly country-level economic growth measures observed for 20 countries over 30 years

It should be clear from our previous discussions that both of these types of grouped data likely violate at least one OLS assumption: the assumption of independence of units. That is, the residual of observation  $k$  does not predict the residual of observation  $n$ , or, in more descriptive terms, observations  $k$  and  $n$  have nothing in common that the regression model is not already accounting for. Grouped data likely violate this assumption because observations that are part of the same group likely share some characteristics that are difficult to model.

Similarly, grouped data may also be likely to produce heteroskedastic errors if groups exhibit different patterns of residuals. Think of the following two examples:

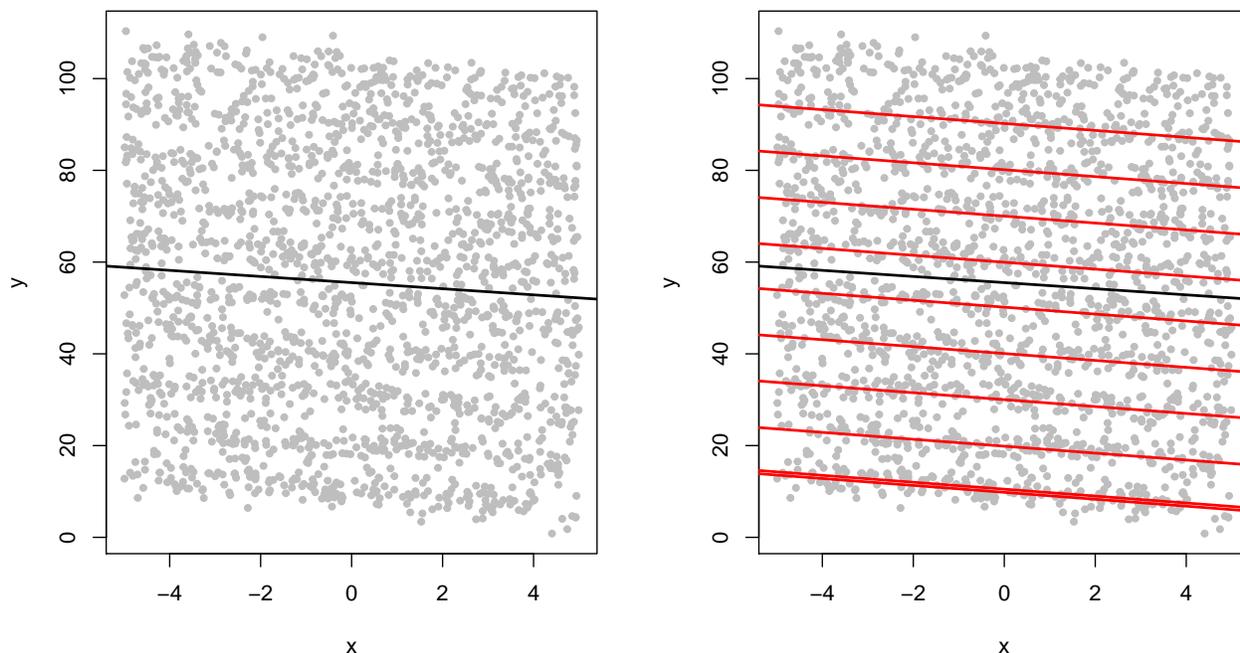
1. In studies of opinions toward policies, individuals in one country may all be framing their opinion in terms of some shared cultural experience germane to that country. Respondents from the same country are unlikely to be independent from each other.
2. Economic growth over time in the United States may be more volatile than in Sweden due to difficult-to-measure factors. Economic growth measures from the United States are not independent from each other.

## A simple solution: group-specific intercepts

From [Day 9](#) of our seminar, you recall that dummy variables can be used to assign separate intercepts to different groups. This means that each group receives a separate intercept for its (virtual) regression line, while the relationships between all other predictors are the same (i.e., parallel regression lines) for all groups. In a simple regression equation, we can include group dummies  $\alpha_j$  as follows, where groups are indexed by  $j$  and individual observations are indexed by  $i$ :

$$y_{ij} = \alpha_j + \beta x_i + \varepsilon_i$$

You can see the group dummies visualized in the figure on the right. You can also see that *not* accounting for the groups will (a) leave much noise around the estimate for the  $\beta$  (the relationship between  $x$  and  $y$ ). It will also (although in this example only slightly) bias the estimate of  $\beta$ ; you can see this by comparing the slope of the black line and the red lines in the right figure. The black line is the slope from a regression that does *not* account for the groups.



```
## NULL
```

Fitting these separate intercepts for groups is commonly known as **fixed effects**. Informally, fixed effects have some advantages:

- FEs account for any group-level sources of variation in the outcome you're trying to explain - sources of variation that are *not* captured in your regression otherwise.
- FEs allow you to estimate the “base level” of your outcome of interest for each group. That is, you can make statements such as “on average, economic growth is higher in the United States than it is in Sweden.”
- FEs are also popular as a tool for causal identification *under certain circumstances* (see., e.g, chapter 5 in *Mostly Harmless Econometrics* by Angrist and Pischke).

## FEs in nested data: a survey example

To illustrate the advantages of fixed effects in nested data (e.g., survey respondents nested in different geographic units), we'll use a cleaned-up and reduced version of the [Cooperative Study of Electoral Systems](#). I provided these data on my website. We'll use these data to estimate a model of satisfaction with democracy. The variables in this dataset are:

Variable	Description
satisdem	Index of satisfaction with democracy, from 1 (dissatisfied) to 4 (satisfied)
married	Married
employed	Full-time employed
union	Union member
voted	Respondent voted in last election
female	Female respondent
age	Age
education	Education
lrself	Left/right self-placement
cname	Name of the respondent's country and year of the last election

To obtain a list of the countries represented in the survey, we can use the `table()` command:

```
cses.dat <- read.csv("http://www.jkarreth.net/files/cses3_reduced.csv")
table(cses.dat$cname)
```

```
##
##      AUSTRALIA (2007)      AUSTRIA (2008)      CANADA (2008)
##              1168              1004              1731
## CZECH REPUBLIC (2010)    DENMARK (2007)    ESTONIA (2011)
##              1510              1273              770
##      FINLAND (2011)      FRANCE (2007)      GERMANY (2009)
##              1178              1940              1784
##      GREECE (2009)      ICELAND (2009)      IRELAND (2007)
##              923              1094              674
##      ISRAEL (2006)      JAPAN (2007)      NETHERLANDS (2010)
##              1065              1098              1897
## NEW ZEALAND (2008)    NORWAY (2009)      POLAND (2007)
##              803              1700              1788
```

```
##          PORTUGAL (2009)          SLOVAKIA (2010)          SLOVENIA (2008)
##                1024                984                682
## SOUTH KOREA (2008)          SPAIN (2008)          SWEDEN (2006)
##                745                936                1019
## SWITZERLAND (2007)          TAIWAN (2008)          UNITED STATES (2008)
##                2945                1024                1830
##          URUGUAY (2009)
##                812
```

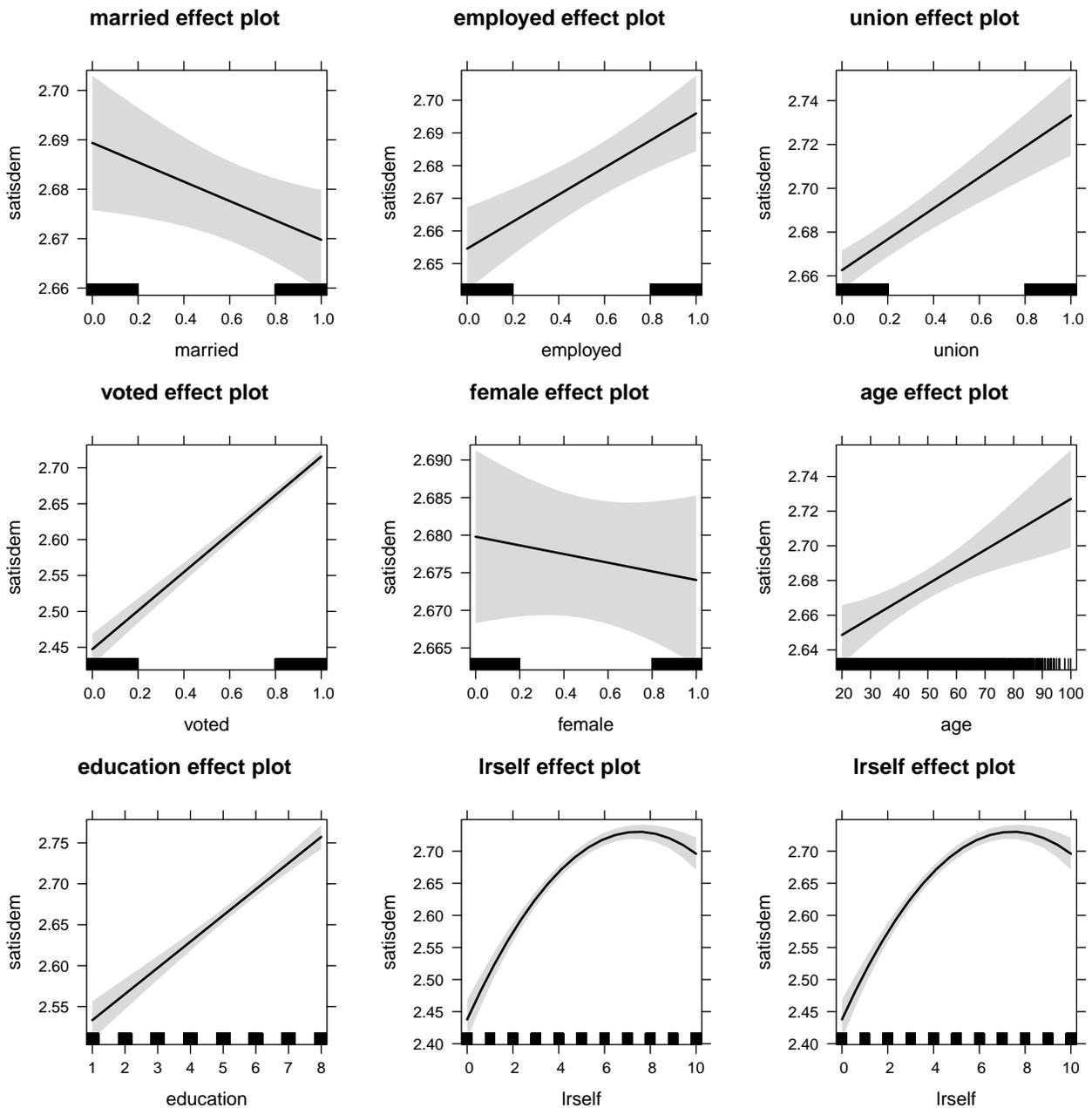
For this tutorial, I'll treat the outcome - satisfaction with democracy, a 4-point scale - as continuous. Estimating a model with individual predictors "pools" all respondents and returns the following results:

```
sd.pooled <- lm(satisdem ~ married + employed + union + voted +
               female + age + education + lrself + I(lrself^2), data = cses.dat)
summary(sd.pooled)
```

```
##
## Call:
## lm(formula = satisdem ~ married + employed + union + voted +
##     female + age + education + lrself + I(lrself^2), data = cses.dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.9506 -0.6481  0.2277  0.3564  1.9264
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.9640444  0.0270099  72.716 < 2e-16 ***
## married      -0.0196254  0.0087782  -2.236 0.025378 *
## employed      0.0413796  0.0092576   4.470 7.85e-06 ***
## union         0.0705390  0.0104848   6.728 1.75e-11 ***
## voted         0.2682937  0.0118711  22.601 < 2e-16 ***
## female       -0.0057596  0.0082341  -0.699 0.484256
## age           0.0009801  0.0002677   3.662 0.000251 ***
## education     0.0319880  0.0024815  12.891 < 2e-16 ***
## lrself        0.0785386  0.0060129  13.062 < 2e-16 ***
## I(lrself^2)  -0.0052712  0.0005484  -9.612 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7631 on 34853 degrees of freedom
## (538 observations deleted due to missingness)
## Multiple R-squared:  0.03281,    Adjusted R-squared:  0.03256
## F-statistic: 131.4 on 9 and 34853 DF,  p-value: < 2.2e-16
```

Sidenote: the "effects" package you encountered last week can also be used for OLS:

```
library(effects)
plot(allEffects(sd.pooled))
```



As you saw above, the countries in this survey are quite diverse. Would it make sense to consider respondents from the same country independent? And would you be worried about pooling respondents from a new democracy such as Slovakia and an established democracy such as the United States? Fixed effects can partially address these concerns.

```
sd.fe <- lm(satisdem ~ married + employed + union + voted +
            female + age + education + lrsel + I(lrsel^2) + factor(cname),
            data = cses.dat)
summary(sd.fe)
```

```
##
## Call:
## lm(formula = satisdem ~ married + employed + union + voted +
```

```

##   female + age + education + lrself + I(lrself^2) + factor(cname),
##   data = cses.dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.32681 -0.34898  0.04048  0.45081  2.24753
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.6234390  0.0337276  77.783 < 2e-16
## married           0.0117220  0.0080508   1.456 0.145400
## employed        -0.0002075  0.0085399  -0.024 0.980618
## union           -0.0033486  0.0108239  -0.309 0.757043
## voted            0.1687882  0.0112828  14.960 < 2e-16
## female          -0.0230978  0.0075162  -3.073 0.002120
## age             -0.0008120  0.0002492  -3.258 0.001123
## education        0.0237596  0.0024210   9.814 < 2e-16
## lrself           0.0660353  0.0055609  11.875 < 2e-16
## I(lrself^2)     -0.0040384  0.0005085  -7.942 2.05e-15
## factor(cname)AUSTRIA (2008) -0.2475184  0.0300769  -8.230 < 2e-16
## factor(cname)CANADA (2008) -0.2258618  0.0266192  -8.485 < 2e-16
## factor(cname)CZECH REPUBLIC (2010) -0.7069347  0.0274955 -25.711 < 2e-16
## factor(cname)DENMARK (2007)  0.1488733  0.0288080   5.168 2.38e-07
## factor(cname)ESTONIA (2011) -0.6335524  0.0325568 -19.460 < 2e-16
## factor(cname)FINLAND (2011) -0.2700551  0.0290803  -9.287 < 2e-16
## factor(cname)FRANCE (2007)  -0.3603989  0.0259710 -13.877 < 2e-16
## factor(cname)GERMANY (2009) -0.4899690  0.0265025 -18.488 < 2e-16
## factor(cname)GREECE (2009) -1.0509429  0.0308224 -34.097 < 2e-16
## factor(cname)ICELAND (2009) -0.8102822  0.0300474 -26.967 < 2e-16
## factor(cname)IRELAND (2007) -0.1718445  0.0338612  -5.075 3.90e-07
## factor(cname)ISRAEL (2006)  -0.9488008  0.0298130 -31.825 < 2e-16
## factor(cname)JAPAN (2007)   -0.6195115  0.0294470 -21.038 < 2e-16
## factor(cname)NETHERLANDS (2010) -0.2640142  0.0260388 -10.139 < 2e-16
## factor(cname)NEW ZEALAND (2008) -0.3921531  0.0321194 -12.209 < 2e-16
## factor(cname)NORWAY (2009)  -0.0091268  0.0265973  -0.343 0.731489
## factor(cname)POLAND (2007)  -0.7727139  0.0269114 -28.713 < 2e-16
## factor(cname)PORTUGAL (2009) -0.8151904  0.0300565 -27.122 < 2e-16
## factor(cname)SLOVAKIA (2010) -0.7804179  0.0303292 -25.732 < 2e-16
## factor(cname)SLOVENIA (2008) -0.8120247  0.0338397 -23.996 < 2e-16
## factor(cname)SOUTH KOREA (2008) -0.7337151  0.0331257 -22.149 < 2e-16
## factor(cname)SPAIN (2008)   -0.1404658  0.0310338  -4.526 6.02e-06
## factor(cname)SWEDEN (2006)  -0.1101505  0.0301230  -3.657 0.000256
## factor(cname)SWITZERLAND (2007) -0.1161815  0.0244739  -4.747 2.07e-06
## factor(cname)TAIWAN (2008)  -0.5626945  0.0299643 -18.779 < 2e-16
## factor(cname)UNITED STATES (2008) 0.0151381  0.0277483   0.546 0.585379
## factor(cname)URUGUAY (2009) -0.0910952  0.0320575  -2.842 0.004491
##
## (Intercept)      ***
## married
## employed
## union
## voted            ***
## female           **
## age              **

```

```

## education ***
## lrself ***
## I(lrself^2) ***
## factor(cname)AUSTRIA (2008) ***
## factor(cname)CANADA (2008) ***
## factor(cname)CZECH REPUBLIC (2010) ***
## factor(cname)DENMARK (2007) ***
## factor(cname)ESTONIA (2011) ***
## factor(cname)FINLAND (2011) ***
## factor(cname)FRANCE (2007) ***
## factor(cname)GERMANY (2009) ***
## factor(cname)GREECE (2009) ***
## factor(cname)ICELAND (2009) ***
## factor(cname)IRELAND (2007) ***
## factor(cname)ISRAEL (2006) ***
## factor(cname)JAPAN (2007) ***
## factor(cname)NETHERLANDS (2010) ***
## factor(cname)NEW ZEALAND (2008) ***
## factor(cname)NORWAY (2009) ***
## factor(cname)POLAND (2007) ***
## factor(cname)PORTUGAL (2009) ***
## factor(cname)SLOVAKIA (2010) ***
## factor(cname)SLOVENIA (2008) ***
## factor(cname)SOUTH KOREA (2008) ***
## factor(cname)SPAIN (2008) ***
## factor(cname)SWEDEN (2006) ***
## factor(cname)SWITZERLAND (2007) ***
## factor(cname)TAIWAN (2008) ***
## factor(cname)UNITED STATES (2008) ***
## factor(cname)URUGUAY (2009) **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6946 on 34826 degrees of freedom
## (538 observations deleted due to missingness)
## Multiple R-squared:  0.1992, Adjusted R-squared:  0.1983
## F-statistic: 240.6 on 36 and 34826 DF, p-value: < 2.2e-16

```

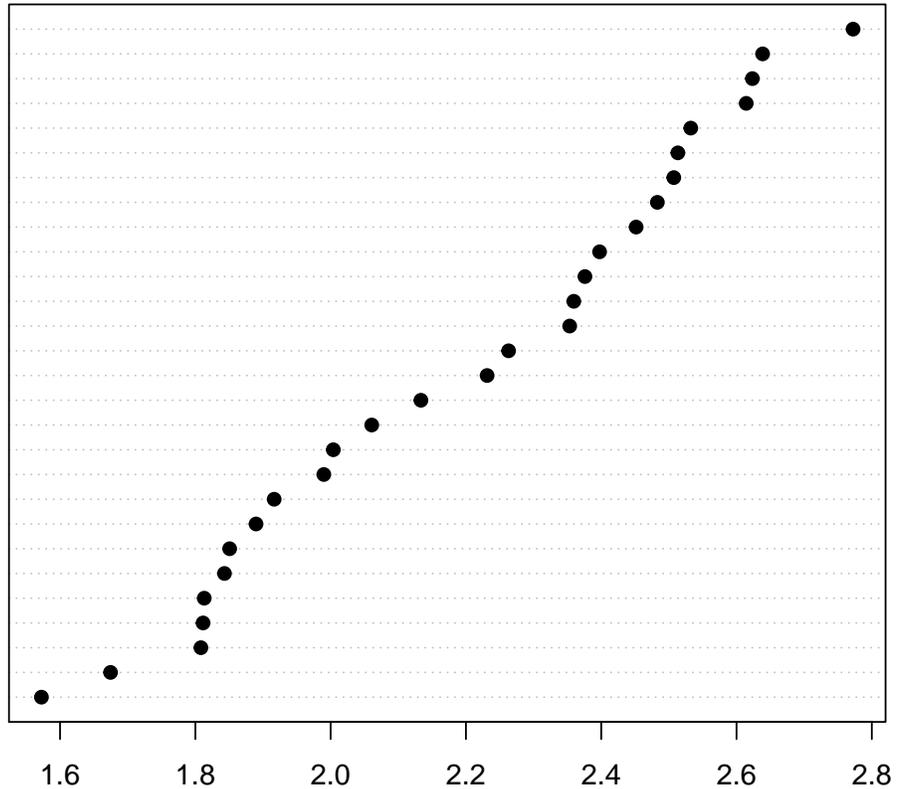
You can visualize the fixed effects using the `dotchart()` function you learned last week:

```

fe.dat <- data.frame(fe = coef(sd.fe)[c(1, grep("cname", names(coef(sd.fe))))],
                    cname = unique(cses.dat$cname))
mean <- fe.dat[1, ]$fe
fe.dat[-1, ]$fe <- fe.dat[-1, ]$fe + mean
fe.dat <- fe.dat[order(fe.dat$fe), ]
dotchart(x = fe.dat$fe, label = fe.dat$cname, pch = 19,
         xlab = "Satisfaction with democracy")

```

DENMARK (2007)  
 UNITED STATES (2008)  
 AUSTRALIA (2007)  
 NORWAY (2009)  
 URUGUAY (2009)  
 SWEDEN (2006)  
 SWITZERLAND (2007)  
 SPAIN (2008)  
 IRELAND (2007)  
 CANADA (2008)  
 AUSTRIA (2008)  
 NETHERLANDS (2010)  
 FINLAND (2011)  
 FRANCE (2007)  
 NEW ZEALAND (2008)  
 GERMANY (2009)  
 TAIWAN (2008)  
 JAPAN (2007)  
 ESTONIA (2011)  
 CZECH REPUBLIC (2010)  
 SOUTH KOREA (2008)  
 POLAND (2007)  
 SLOVAKIA (2010)  
 ICELAND (2009)  
 SLOVENIA (2008)  
 PORTUGAL (2009)  
 ISRAEL (2006)  
 GREECE (2009)



Satisfaction with democracy

Using fixed effects here helps you account for unobserved sources of variation at the country-level. While this may be desirable, it also limits you: all country-level sources of explanation for satisfaction with democracy are now soaked up in the fixed effects. That is, if you would like to estimate the relationship between country-level covariates (such as the state of the economy, the integrity of the democratic process, etc.), you cannot do so while fixed effects are in the model. This should be logical: a dummy variable for each country and any country-level predictor will be perfectly collinear and the model is not identified.

See for yourself: here, I include a country-level predictor for economic growth before the election, `gdpgrowth`:

```
sd.fe2 <- lm(satisdem ~ married + employed + union + voted +
            female + age + education + lrself + I(lrself^2) + factor(cname) +
            gdpgrowth,
            data = cses.dat)
summary(sd.fe2)
```

```
##
## Call:
## lm(formula = satisdem ~ married + employed + union + voted +
##     female + age + education + lrself + I(lrself^2) + factor(cname) +
##     gdpgrowth, data = cses.dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.32681 -0.34898  0.04048  0.45081  2.24753
##
## Coefficients: (1 not defined because of singularities)
```

	Estimate	Std. Error	t value	Pr(> t )
## (Intercept)	2.6234390	0.0337276	77.783	< 2e-16
## married	0.0117220	0.0080508	1.456	0.145400
## employed	-0.0002075	0.0085399	-0.024	0.980618
## union	-0.0033486	0.0108239	-0.309	0.757043
## voted	0.1687882	0.0112828	14.960	< 2e-16
## female	-0.0230978	0.0075162	-3.073	0.002120
## age	-0.0008120	0.0002492	-3.258	0.001123
## education	0.0237596	0.0024210	9.814	< 2e-16
## lrself	0.0660353	0.0055609	11.875	< 2e-16
## I(lrself^2)	-0.0040384	0.0005085	-7.942	2.05e-15
## factor(cname)AUSTRIA (2008)	-0.2475184	0.0300769	-8.230	< 2e-16
## factor(cname)CANADA (2008)	-0.2258618	0.0266192	-8.485	< 2e-16
## factor(cname)CZECH REPUBLIC (2010)	-0.7069347	0.0274955	-25.711	< 2e-16
## factor(cname)DENMARK (2007)	0.1488733	0.0288080	5.168	2.38e-07
## factor(cname)ESTONIA (2011)	-0.6335524	0.0325568	-19.460	< 2e-16
## factor(cname)FINLAND (2011)	-0.2700551	0.0290803	-9.287	< 2e-16
## factor(cname)FRANCE (2007)	-0.3603989	0.0259710	-13.877	< 2e-16
## factor(cname)GERMANY (2009)	-0.4899690	0.0265025	-18.488	< 2e-16
## factor(cname)GREECE (2009)	-1.0509429	0.0308224	-34.097	< 2e-16
## factor(cname)ICELAND (2009)	-0.8102822	0.0300474	-26.967	< 2e-16
## factor(cname)IRELAND (2007)	-0.1718445	0.0338612	-5.075	3.90e-07
## factor(cname)ISRAEL (2006)	-0.9488008	0.0298130	-31.825	< 2e-16
## factor(cname)JAPAN (2007)	-0.6195115	0.0294470	-21.038	< 2e-16
## factor(cname)NETHERLANDS (2010)	-0.2640142	0.0260388	-10.139	< 2e-16
## factor(cname)NEW ZEALAND (2008)	-0.3921531	0.0321194	-12.209	< 2e-16
## factor(cname)NORWAY (2009)	-0.0091268	0.0265973	-0.343	0.731489
## factor(cname)POLAND (2007)	-0.7727139	0.0269114	-28.713	< 2e-16
## factor(cname)PORTUGAL (2009)	-0.8151904	0.0300565	-27.122	< 2e-16
## factor(cname)SLOVAKIA (2010)	-0.7804179	0.0303292	-25.732	< 2e-16
## factor(cname)SLOVENIA (2008)	-0.8120247	0.0338397	-23.996	< 2e-16
## factor(cname)SOUTH KOREA (2008)	-0.7337151	0.0331257	-22.149	< 2e-16
## factor(cname)SPAIN (2008)	-0.1404658	0.0310338	-4.526	6.02e-06
## factor(cname)SWEDEN (2006)	-0.1101505	0.0301230	-3.657	0.000256
## factor(cname)SWITZERLAND (2007)	-0.1161815	0.0244739	-4.747	2.07e-06
## factor(cname)TAIWAN (2008)	-0.5626945	0.0299643	-18.779	< 2e-16
## factor(cname)UNITED STATES (2008)	0.0151381	0.0277483	0.546	0.585379
## factor(cname)URUGUAY (2009)	-0.0910952	0.0320575	-2.842	0.004491
## gdpgrowth	NA	NA	NA	NA
## (Intercept)	***			
## married				
## employed				
## union				
## voted	***			
## female	**			
## age	**			
## education	***			
## lrself	***			
## I(lrself^2)	***			
## factor(cname)AUSTRIA (2008)	***			
## factor(cname)CANADA (2008)	***			
## factor(cname)CZECH REPUBLIC (2010)	***			
## factor(cname)DENMARK (2007)	***			

```

## factor(cname)ESTONIA (2011)      ***
## factor(cname)FINLAND (2011)     ***
## factor(cname)FRANCE (2007)      ***
## factor(cname)GERMANY (2009)     ***
## factor(cname)GREECE (2009)     ***
## factor(cname)ICELAND (2009)     ***
## factor(cname)IRELAND (2007)     ***
## factor(cname)ISRAEL (2006)      ***
## factor(cname)JAPAN (2007)       ***
## factor(cname)NETHERLANDS (2010) ***
## factor(cname)NEW ZEALAND (2008) ***
## factor(cname)NORWAY (2009)      ***
## factor(cname)POLAND (2007)      ***
## factor(cname)PORTUGAL (2009)    ***
## factor(cname)SLOVAKIA (2010)    ***
## factor(cname)SLOVENIA (2008)    ***
## factor(cname)SOUTH KOREA (2008) ***
## factor(cname)SPAIN (2008)       ***
## factor(cname)SWEDEN (2006)      ***
## factor(cname)SWITZERLAND (2007) ***
## factor(cname)TAIWAN (2008)      ***
## factor(cname)UNITED STATES (2008)
## factor(cname)URUGUAY (2009)     **
## gdpgrowth
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6946 on 34826 degrees of freedom
## (538 observations deleted due to missingness)
## Multiple R-squared:  0.1992, Adjusted R-squared:  0.1983
## F-statistic: 240.6 on 36 and 34826 DF,  p-value: < 2.2e-16

```

The coefficient for GDP growth cannot be estimated because GDP growth does not vary for all respondents within one country.

## FEs in TSCS data: a political economy example

You can also think of time-series cross-sectional data as nested: observations are “nested” in time. As briefly discussed above, it may be desirable or necessary to account for unobserved sources of variation by using fixed effects. An example comes from Geoffrey Garrett’s 1998 book *Partisan Politics in the Global Economy*. These data are discussed in section 5.2 of a recent article by Nathaniel Beck and Jonathan Katz, “[Modeling Dynamics in Time-Series–Cross-Section Political Economy Data](#)”, in the *Annual Review of Political Science* (Vol. 14: pp. 331-352). One of the chapters in Garrett’s book asks left-leaning governments in OECD countries are associated with lower economic growth rates. To answer this question, Garrett regresses economic growth rates on a measure of leftist participation in government (the percentage of cabinet posts held by members of Left parties), an indicator of the global economic climate (OECD demand), an indicator for oil dependency of a country, and an indicator for the institutionalization of centralized wage bargaining or corporatism. Garrett’s argument suggests that leftist governments should oversee less economic growth only when wage bargaining is not centralized, therefore he includes an interaction term of leftist government participation and corporatism.

Replication data are available from Jonathan Katz’s [dataverse page](#) for the 2011 article, as a tab-delimited file under the name “garrett1998.tab”. You can download the file from there. The relevant variables in it are:

Variable	Description
year	Year of the observation
country	Country (as <a href="#">Correlates of War country codes</a> )
gdp	Growth rate of GDP
gdpl	Growth rate of GDP (lagged from previous year)
leftlab	% of cabinet members from Left parties (transformed)
corp	Corporatism index
demand	Overall OECD GDP growth, weighted for each country by its trade with the other OECD nations
oild	Oil dependence of the economy

First, I load the data and assign country names to the country codes. For this, I use the “countrycode” package.

```
garrett.dat <- read.table("http://www.jkarreth.net/files/garrett1998.tab", header = TRUE)
summary(garrett.dat)
```

```
##      country      year      unem      infl
## Min.   : 2.0    Min.   :1966  Min.   : 0.6848  Min.   : -0.700
## 1st Qu.:210.0  1st Qu.:1972  1st Qu.: 2.0992  1st Qu.: 3.700
## Median :282.5  Median :1978  Median : 4.5000  Median : 5.900
## Mean   :287.4  Mean   :1978  Mean   : 4.9939  Mean   : 6.688
## 3rd Qu.:380.0  3rd Qu.:1984  3rd Qu.: 7.3000  3rd Qu.: 9.075
## Max.   :740.0  Max.   :1990  Max.   :13.0000  Max.   :24.500
##      gdp      uneml      infl1      gdpl
## Min.   : -4.300  Min.   : 0.600  Min.   : -0.700  Min.   : -4.300
## 1st Qu.: 1.877  1st Qu.: 2.000  1st Qu.: 3.800  1st Qu.: 2.000
## Median : 3.200  Median : 3.900  Median : 5.900  Median : 3.300
## Mean   : 3.254  Mean   : 4.832  Mean   : 6.683  Mean   : 3.336
## 3rd Qu.: 4.700  3rd Qu.: 7.100  3rd Qu.: 8.900  3rd Qu.: 4.795
## Max.   :12.800  Max.   :13.000  Max.   :24.500  Max.   :12.800
##      trade      capmob      oild      Icc_2
## Min.   : 9.623  Min.   :0.0000  Min.   : -0.117810  Min.   :0.00000
## 1st Qu.: 41.419  1st Qu.:0.0000  1st Qu.: 0.003662  1st Qu.:0.00000
## Median : 52.624  Median :1.0000  Median : 0.013674  Median :0.00000
## Mean   : 57.076  Mean   :0.8914  Mean   : 0.015280  Mean   :0.07143
## 3rd Qu.: 71.847  3rd Qu.:1.0000  3rd Qu.: 0.030052  3rd Qu.:0.00000
## Max.   :146.020  Max.   :4.0000  Max.   : 0.083460  Max.   :1.00000
##      Icc_3      Icc_4      Icc_5      Icc_6
## Min.   :0.00000  Min.   :0.00000  Min.   :0.00000  Min.   :0.00000
## 1st Qu.:0.00000  1st Qu.:0.00000  1st Qu.:0.00000  1st Qu.:0.00000
## Median :0.00000  Median :0.00000  Median :0.00000  Median :0.00000
## Mean   :0.07143  Mean   :0.07143  Mean   :0.07143  Mean   :0.07143
## 3rd Qu.:0.00000  3rd Qu.:0.00000  3rd Qu.:0.00000  3rd Qu.:0.00000
## Max.   :1.00000  Max.   :1.00000  Max.   :1.00000  Max.   :1.00000
##      Icc_7      Icc_8      Icc_9      Icc_10
## Min.   :0.00000  Min.   :0.00000  Min.   :0.00000  Min.   :0.00000
## 1st Qu.:0.00000  1st Qu.:0.00000  1st Qu.:0.00000  1st Qu.:0.00000
## Median :0.00000  Median :0.00000  Median :0.00000  Median :0.00000
## Mean   :0.07143  Mean   :0.07143  Mean   :0.07143  Mean   :0.07143
## 3rd Qu.:0.00000  3rd Qu.:0.00000  3rd Qu.:0.00000  3rd Qu.:0.00000
## Max.   :1.00000  Max.   :1.00000  Max.   :1.00000  Max.   :1.00000
##      Icc_11      Icc_12      Icc_13      Icc_14
```

```

## Min. :0.00000 Min. :0.00000 Min. :0.00000 Min. :0.00000
## 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:0.00000
## Median :0.00000 Median :0.00000 Median :0.00000 Median :0.00000
## Mean :0.07143 Mean :0.07143 Mean :0.07143 Mean :0.07143
## 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:0.00000
## Max. :1.00000 Max. :1.00000 Max. :1.00000 Max. :1.00000
## per6673 per7479 per8084 per8690 corp
## Min. :0.00 Min. :0.00 Min. :0.0 Min. :0.0 Min. :0.415
## 1st Qu.:0.00 1st Qu.:0.00 1st Qu.:0.0 1st Qu.:0.0 1st Qu.:2.189
## Median :0.00 Median :0.00 Median :0.0 Median :0.0 Median :3.230
## Mean :0.32 Mean :0.24 Mean :0.2 Mean :0.2 Mean :3.016
## 3rd Qu.:1.00 3rd Qu.:0.00 3rd Qu.:0.0 3rd Qu.:0.0 3rd Qu.:3.794
## Max. :1.00 Max. :1.00 Max. :1.0 Max. :1.0 Max. :4.820
## leftlab clint demand
## Min. :0.05066 Min. : 0.085 Min. : -39.17
## 1st Qu.:1.14301 1st Qu.: 2.542 1st Qu.: 92.59
## Median :2.17484 Median : 6.396 Median :177.52
## Mean :2.02770 Mean : 6.763 Mean :187.42
## 3rd Qu.:2.80958 3rd Qu.:10.015 3rd Qu.:257.06
## Max. :3.57084 Max. :17.162 Max. :644.79

```

```

# install.packages("countrycode")
library(countrycode)
garrett.dat$name <- countrycode(sourcevar = garrett.dat$country,
                               origin = "cown",
                               destination = "country.name")
table(garrett.dat$name)

```

```

##
## Austria Belgium
## 25 25
## Canada Denmark
## 25 25
## Federal Republic of Germany Finland
## 25 25
## France Italy
## 25 25
## Japan Netherlands
## 25 25
## Norway Sweden
## 25 25
## United Kingdom United States
## 25 25

```

```

garrett.dat$name <- ifelse(garrett.dat$name == "Federal Republic of Germany",
                          "Germany",
                          garrett.dat$name)
table(garrett.dat$name)

```

```

##
## Austria Belgium Canada Denmark Finland
## 25 25 25 25 25
## France Germany Italy Japan Netherlands

```

```
##           25           25           25           25           25
##      Norway      Sweden United Kingdom United States
##           25           25           25           25
```

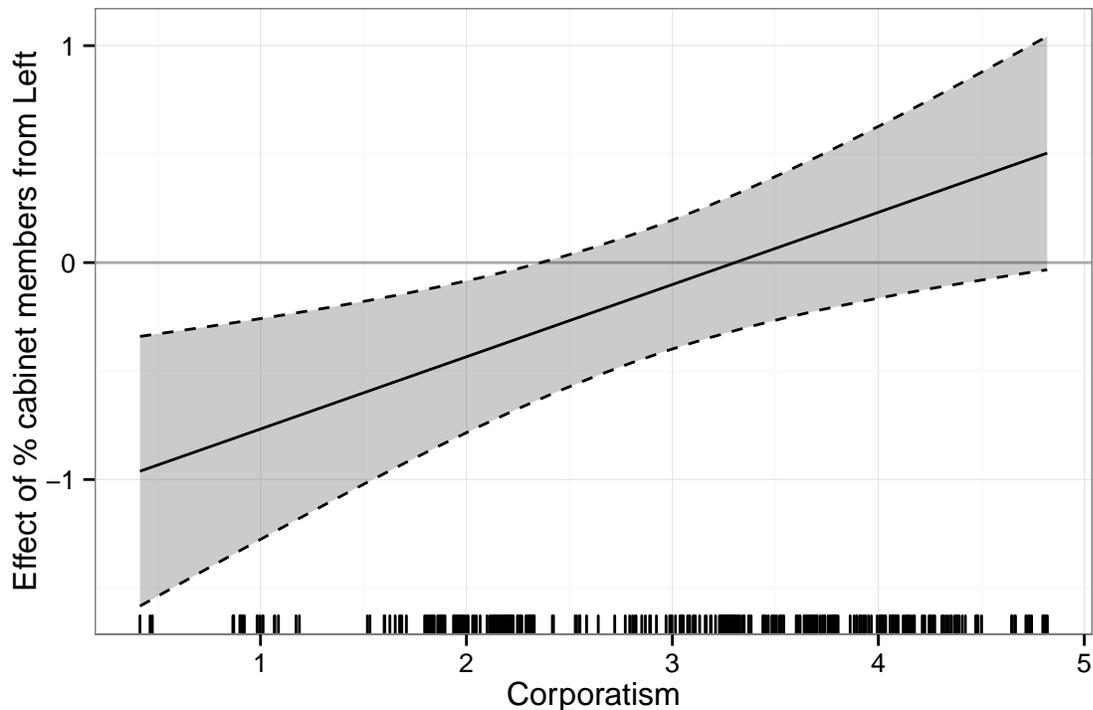
To capture Garrett's argument about that the impact of leftist government participation on economic growth is conditional on the degree of corporatism, I include an interaction term between the two variables in the model below. This model pools all observations and assumes that each observation is independent from each other (once we've accounted for temporal autocorrelation by including a lag of the outcome variable). In other words, there is nothing inherently different about the United States in 1980 compared to Sweden in 1980, once we've controlled for economic growth in the previous year, the share of leftists in the cabinet, the level of corporatism, oil dependence, and the global economic climate. Compare this to Table 5.3 in Garrett (1998).

```
gdp.pooled <- lm(gdp ~ gdpl + oild + demand + leftlab * corp, data = garrett.dat)
summary(gdp.pooled)
```

```
##
## Call:
## lm(formula = gdp ~ gdpl + oild + demand + leftlab * corp, data = garrett.dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -7.1171 -1.2909 -0.0205  1.3530  6.6253
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.7514261  0.7356121   5.100 5.64e-07 ***
## gdpl         0.3696199  0.0478508   7.724 1.25e-13 ***
## oild        -9.3066684  4.2953481  -2.167  0.03095 *
## demand       0.0052377  0.0009239   5.669 3.04e-08 ***
## leftlab     -1.1000959  0.3588064  -3.066  0.00234 **
## corp        -0.8592037  0.2836856  -3.029  0.00264 **
## leftlab:corp  0.3329478  0.1151210   2.892  0.00407 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.099 on 343 degrees of freedom
## Multiple R-squared:  0.2683, Adjusted R-squared:  0.2555
## F-statistic: 20.96 on 6 and 343 DF,  p-value: < 2.2e-16
```

Because the original paper contains an interaction term, we can plot the marginal effects:

```
library(devtools)
source_url("https://raw.githubusercontent.com/jkarreth/JKmisc/master/ggintfun.R")
ggintfun(gdp.pooled, varnames = c("leftlab", "corp"),
         varlabs = c("% cabinet members from Left", "Corporatism"),
         title = FALSE, rug = TRUE,
         twoways = FALSE)
```



However, you may suspect that the control variables listed above do not capture all potential country-level characteristics that might influence economic growth. To soak up all between-country variation, you can include fixed effects and re-estimate the model.

```
gdp.fe <- lm(gdp ~ gdpl + oild + demand + leftlab * corp + factor(cname), data = garrett.dat)
summary(gdp.fe)
```

```
##
## Call:
## lm(formula = gdp ~ gdpl + oild + demand + leftlab * corp + factor(cname),
##     data = garrett.dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6.9246 -1.1679 -0.0681  1.1914  5.4880
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.212508   2.826633   0.075  0.94012
## gdpl           0.250525   0.048714   5.143 4.65e-07 ***
## oild          -18.415379   5.898115  -3.122  0.00195 **
## demand         0.008451   0.001088   7.767 1.02e-13 ***
## leftlab       -0.887424   0.434177  -2.044  0.04175 *
## corp          -0.416717   0.642045  -0.649  0.51676
## factor(cname)Belgium  0.406168   1.173264   0.346  0.72942
## factor(cname)Canada  2.079316   1.772377   1.173  0.24157
## factor(cname)Denmark  0.482628   0.798703   0.604  0.54608
## factor(cname)Finland  1.696598   0.879336   1.929  0.05454 .
## factor(cname)France  2.726026   2.344243   1.163  0.24573
## factor(cname)Germany  1.472264   0.974398   1.511  0.13176
## factor(cname)Italy   2.135334   1.145047   1.865  0.06309 .
```

```

## factor(cname)Japan          4.840057  1.739499  2.782  0.00571 **
## factor(cname)Netherlands    0.554957  1.640841  0.338  0.73542
## factor(cname)Norway         -0.078887  0.975611 -0.081  0.93560
## factor(cname)Sweden         -0.021682  0.637137 -0.034  0.97287
## factor(cname)United Kingdom  1.018206  1.429937  0.712  0.47693
## factor(cname)United States   2.602201  1.843101  1.412  0.15893
## leftlab:corp                0.375131  0.155862  2.407  0.01664 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.966 on 330 degrees of freedom
## Multiple R-squared:  0.3821, Adjusted R-squared:  0.3466
## F-statistic: 10.74 on 19 and 330 DF,  p-value: < 2.2e-16

```

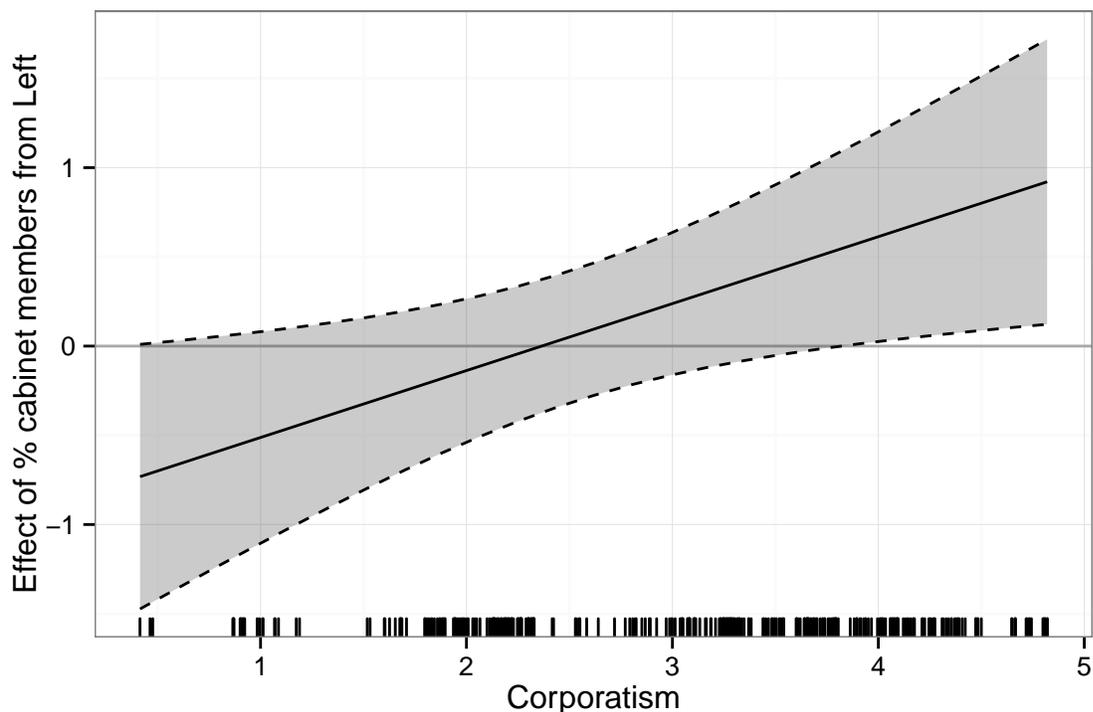
You will notice that the  $R^2$  value has increased (which should make sense). You can now make the statement that all between country-level variation has been taken up by the fixed effects. This has one important implication that's often overlooked. In this model with fixed effects, the coefficient estimates on leftist governments, corporatism, global demand, and oil dependence now **take on a fundamentally different meaning: they express the relationship between a within-country change on  $x$  and  $y$** . That is, you can **not** interpret the coefficient on oil dependency as “countries that depend more on oil exhibit less economic growth compared to other less oil-dependent countries.” Rather, these coefficients mean that “as a country becomes more dependent on oil, it exhibits lower economic growth compared to the same country when it was less dependent on oil.” This is an important distinction for when it comes to testing hypotheses about between-country variation: fixed effects do not allow this interpretation.

Lastly, you can plot the marginal effects again:

```

ggintfun(gdp.fe, varnames = c("leftlab", "corp"),
  varlabs = c("% cabinet members from Left", "Corporatism"),
  title = FALSE, rug = TRUE,
  twoways = FALSE)

```



You will notice that once between-country variance is taken up by the inclusion of fixed effects, the following interpretation emerges: when a country moves toward more centralized wage bargaining, leftist cabinet participation is associated with more economic growth, while leftist cabinet participation has no impact on economic growth in the same country with lower levels of centralized wage bargaining.

## Fixed effects: some caveats

As you saw in the discussion of the two examples above, fixed effects can be useful to account for unobserved confounders at the group level in grouped data. However, using fixed effects in regression has some consequences and comes with some caveats. Some of these are:

- The interpretation of regression coefficients changes *fundamentally* and does not allow for the testing of hypotheses about differences *between* groups (see the Garrett example)
- Fixed effects take up *all* variance between groups. This means that any covariates that are constant at the group level (e.g., the geographic size of a country) cannot be estimated in fixed-effects models
- Fixed effects estimate residual sources of between-group variation, but do not help you in *explaining* these sources of variation.
- Other important caveats can be found in the articles by [Clark and Linzer](#) and [Bell and Jones](#) and the references therein.

*This discussion barely scratches the surface of grouped data and fixed effects, and does not provide any theoretical foundation for their use. Please see the recommendations above for further and important background on this topic.*