SOC542 Statistical Methods in Sociology II

Further directions in statistics: Structure and causality

Thomas Davidson

Rutgers University

April 24, 2023

Course updates

- Updates due Wednesday 4/26 at 5pm via email
- Presentations next week in class
  - 10 minutes to present final project
    - Introduction
    - Data
    - Methodology
    - Main results
    - Robustness check(s)
    - Conclusions

Plan

- Violations of regression assumptions
- Robust and clustered standard errors
- Fixed effects
- Random effects
- Space, time and social structure
- A brief intro to causal inference and regression

Violations of regression assumptions

IID and heteroskedasticity

- Our approach to regression modeling has been based on the assumption that our data are independently and identically distributed
  - e.g. Random samples from a known population
- In practice, this assumption is often violated
  - Groups with different distributions
  - Non-independent observations
- OLS assumes that residuals are homoskedastic, but observed data often have heteroskedastic structures.
  - This is particularly common if data are sampled from different groups with variation in the underlying data generation process.

Violations of regression assumptions

Impact on standard errors

- Confidence intervals that are too narrow too narrow
- More likely to commit Type I errors (false positives) by incorrectly rejecting the null hypothesis
- Inaccurate description of a plausible range of effect sizes

Violations of regression assumptions

Robust and clustered standard errors

- Adjust standard errors to account for violations of assumptions
  - **Robust/Heteroskedasticity consistent** standard errors
  - **Clustered standard errors** can be used to account for particular types of grouping

Robust and clustered standard errors

Intuition

- Variance component of the model is *inconsistent* due to heteroskedasticity or other model misspecification
  - This implies that we will not converge on the true population parameter, even with large samples.
- Corrections can be applied to variance components using a **sandwich** estimator.[1]

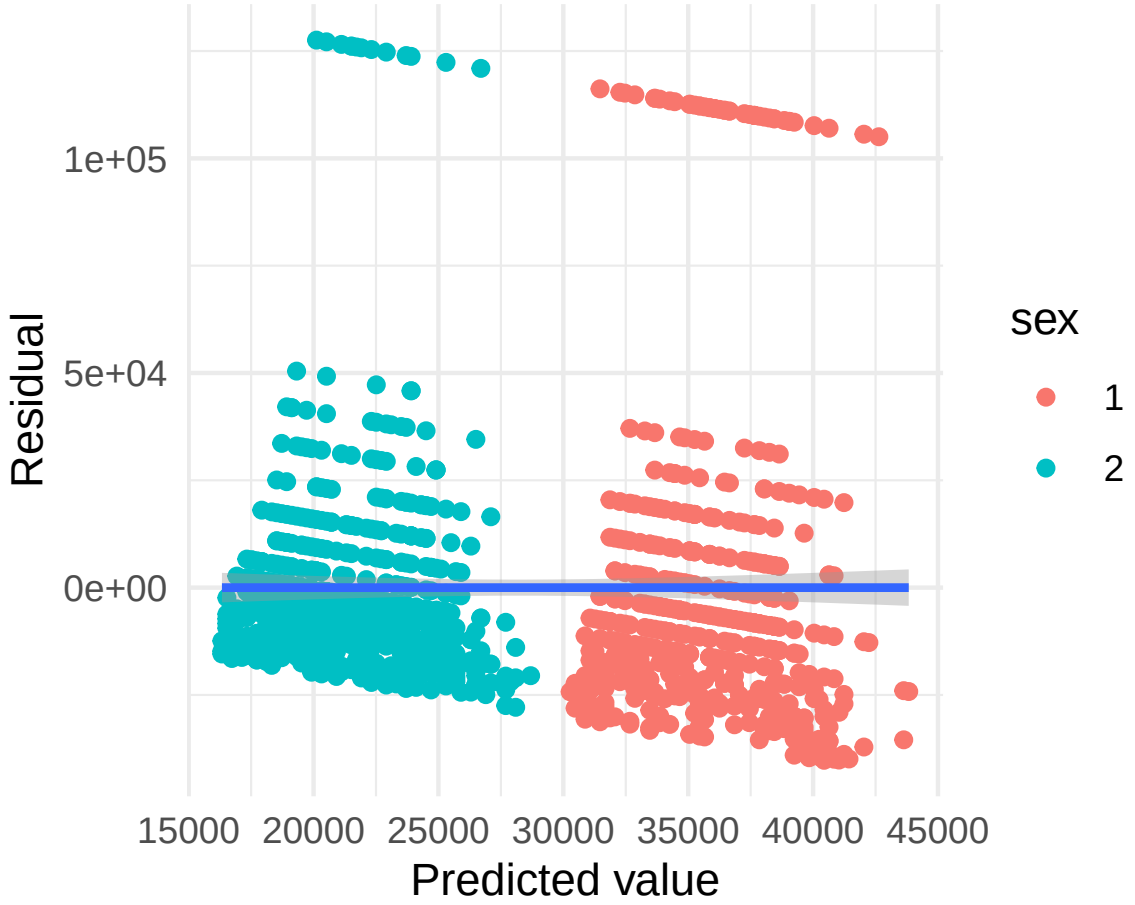
[1] See King and Roberts 2015 for further technical discussion.

Robust and clustered standard errors

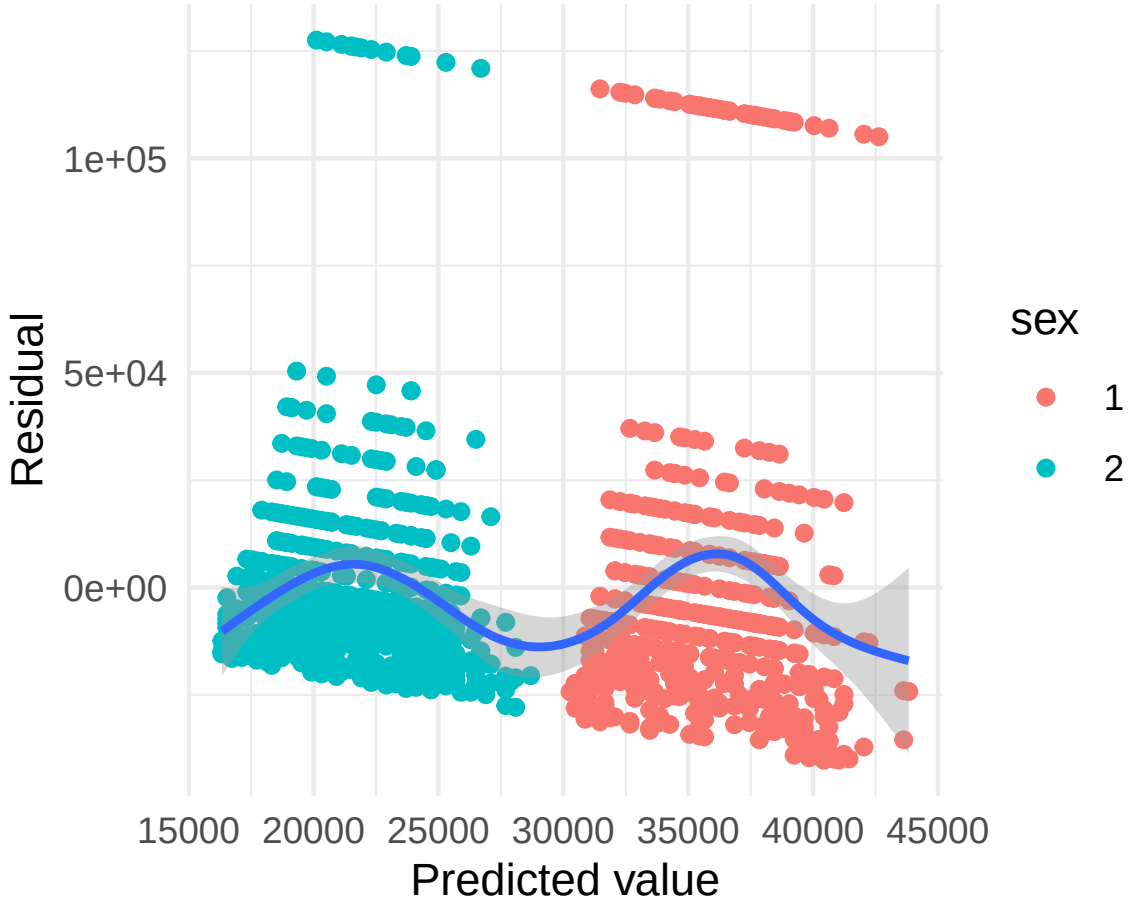
Estimating a simple model: Income as a function of age and sex (GSS 2020)

```
m <- lm(realrinc ~ age + sex, data = gss2020)
```

Heteroskedastic residuals



Heteroskedastic residuals



Robust and clustered standard errors

Calculation in R

There is no need to re-estimate the model. Robust standard errors can be calculated using `sandwich::vcocHC`. The `lmtest:coeftest` function allows us to easily apply the function and format the adjusted model for presentation.[2]

```
library(sandwich)
library(lmtest)
m.r <- coeftest(m, vcov = vcovHC)
```

[2] Grant McDermott's blog has an excellent walkthrough of standard error adjustments using this function.

Robust and clustered standard errors

Clustering by sex

We can use the same function to apply other kinds of standard error correction. For example, we could cluster the errors by sex (although this is not warranted in this case).

```
m.r.g <- coeftest(m, vcov = vcovCL(m, cluster = ~ sex))
```

Robust and clustered standard errors

	OLS	OLS (robust)	OLS (clustered)
(Intercept)	26285.153***	26285.153***	26285.153***
	(3458.317)	(3063.128)	(1929.282)
age	199.354**	199.354**	199.354***
	(66.361)	(61.976)	(40.463)
sex2	-13940.750***	-13940.750***	-13940.750***
	(1915.327)	(1961.312)	(68.045)
Num.Obs.	1077	1077	1077
R2	0.057		
R2 Adj.	0.055		
F	32.385		

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

Robust and clustered standard errors

Caveats

- Robust and clustered standard errors have become popular and are often the default approach in applied econometrics
  - Stata makes it particularly easy to specify them: `reg x y, robust`
- But standard error corrections are not a panacea and do not address underlying issues with model misspecification, as King and Roberts (2015) demonstrate.

Fixed effects

- **Fixed effects** are a useful tool for dealing with unobservables and reducing the threat of omitted variable bias when data have a grouping structure.
  - We previously used fixed-effects to account for any unexplained village-level factors when using the Diffusion of Microfinance dataset.

Fixed effects

- A fixed-effects model can be written like a standard regression model. γ is a vector of coefficients, one dummy variable for each group.

$$y_i = \beta_1 x_i + \gamma_j + u_i$$

- The γ_j term soaks up any within-group variation.
- It is common to drop the intercept from these models, although it is not required.

Fixed effects

Pooling

- **Pooling** refers to how observations are pooled together to estimate averages.
- Considering data with a grouped structure,
  - Standard regression approaches imply **complete pooling** since all available data to estimate a population mean.
    - Any variation between groups is effectively ignored.
  - Fixed-effects regression implies **no pooling** as a separate mean is estimated for each group.
    - No information is shared across groups. Assumption that variation between groups is effectively infinite.

Data and Methodology

- GSS panel data
  - Sample of 2016 and 2018 respondents were re-interviewed in 2020 (online)
  - Formatted data so each observation is one person-year

Data and Methodology

- Outcome
  - `natcrime`: Are we spending too much, about right, or too little on halting the rising crime rate?
  - Dichotomized (1 = too little, 0 = too much/about right)
- Predictors
  - Sex, age, race, political ideology
- Fixed effects
  - Survey year
  - Region
- Model
  - Linear probability model, listwise deletion to drop missing observations

Fixed effects

Implementation in R

We can easily specify fixed effects models using the `fixest` package.[3]

```
library(fixest)
ols <- lm(natcrime ~ sex + race + age + polviews, data = gss.new)
fe.r <- feols(natcrime ~ sex + race + age + polviews | region, data = g
fe.y <- feols(natcrime ~ sex + race + age + polviews | year, data = gss
fe.ry <- feols(natcrime ~ sex + race + age + polviews | region + year,
```

[3] By default this model removes the main intercept from models with fixed effects.

Fixed effects

	Pooled	Region FE	Year FE	Both FE
(Intercept)	0.379***			
	(0.060)			
sex2	0.137***	0.147*	0.139	0.149*
	(0.032)	(0.052)	(0.039)	(0.052)
race2	0.143**	0.117	0.140	0.113
	(0.046)	(0.057)	(0.037)	(0.057)
race3	0.190***	0.179**	0.189	0.177**
	(0.057)	(0.044)	(0.062)	(0.042)
age	0.004**	0.004*	0.004	0.004*
	(0.001)	(0.002)	(0.001)	(0.002)
polviewsConservative	0.077	0.078*	0.076	0.077*
	(0.045)	(0.032)	(0.018)	(0.032)
polviewsLiberal	-0.119**	-0.104	-0.119*	-0.104
	(0.038)	(0.046)	(0.016)	(0.046)
Num.Obs.	855	855	855	855
R2	0.070	0.090	0.071	0.092

Fixed effects

Interpretation

- The fixed effects have accounted for unexplained variation between regions and over time, allowing us to measure the aggregate effect of our predictors on the dependent variable.
- The model with region and time is known as a *two-way FE* estimator (TWFE)

Fixed effects

Limitations of fixed effects

- No pooling
  - No information sharing across groups, only within-group variation analyzed
- Perfect multicollinearity
  - Time-invariant group-level variables are perfectly correlated with fixed effects and dropped from the model

Random effects

Comparing fixed and random effects

- Consider case where we observe random variables y and x, where observations belong to j groups.
- The fixed-effects formulation is given by

$$y_i = \beta_1 x_i + \gamma_j + u_i, u_i \sim N(0, \sigma_u^2)$$

- A random-intercepts model takes a more complex formula, where each element of γ_j is drawn from a distribution:

$$y_i = \beta_0 + \beta_1 x_i + \gamma_j + u_i$$

$$u_i \sim N(0, \sigma_u^2)$$

$$\gamma_j \sim N(0, \sigma_\gamma^2)$$

Random effects

Partial pooling and shrinkage

- The RE model considers the groups as related through a common distribution, whereas the entities in an FE model are unconnected.
- Random effects models are characterized by **partial pooling**
  - Information is shared among groups as intercepts are drawn from a common distribution.
- This tends to reduce overfitting compared to no pooling, since information in each group helps to improve estimates for every other group.
- **Shrinkage** describes how group-level estimates are pushed towards a common mean.
  - This is particularly helpful if there are small groups, where group means might be inaccurately estimated with a fixed effects model.

Random effects

Nesting

- Random effects models allow us to directly model more complex nested data structures
  - e.g. Education researhers might want to consider Level 1 (student), Level 2 (classroom), Level 3 (school), Level 4 (district)
- Unlike fixed effects, where all variance is explained by the fixed effect, variables can be incorporated at different levels
- Shrinkage/partial pooling helps to prevent overfitting

Random effects

A note on terminology

- These models are referred to using a range of different names including mixed effects, random effects, and hierarchical models. Moreover, the term "fixed effects" is also used in different ways, adding to the confusion.
- The "fixed part" or "population" component of a random effects model is the part that does not vary across groups.
  - e.g. $y_i = \beta_0 + \beta_1 x_i$
- The "random part" varies across groups
  - e.g. γ_i

Random effects

Estimation in R

The `lme4` package can be used to estimate Maximum Likelihood random effects models in R. `lmer` function can fit a standard model; `glmer` generalizes to other link functions. The random part is specified in parentheses.

```
library(lme4)
re.r <- lmer(natcrime ~ sex + race + age + polviews + (1|region),
             data = gss.new)
re.r.logit <- glmer(natcrime ~ sex + race + age + polviews + (1|region)
                    data = gss.new, family = binomial)
```

Random effects

Estimation in R

We could also allow each respondent to have their own intercept. This kind of model would not be possible if we used fixed-effects.[4]

```
re.r.id.logit <- glmer(natcrime ~ sex + race + age + polviews +
                           (1 |region) + (1 | id),
                       data = gss.new, family = binomial)
```

[4] In this case, it is probably unnecessary since we only have a couple of observations for each respondent.

Random effects

	Region FE	Region RE	Logit	Logit + Resp
sex2	0.147	0.143	1.928	2.238
	(0.052)	(0.033)	(0.297)	(0.481)
race2	0.117	0.130	1.920	2.399
	(0.057)	(0.047)	(0.461)	(0.794)
race3	0.179	0.188	2.685	3.581
	(0.044)	(0.057)	(0.838)	(1.476)
age	0.004	0.004	1.017	1.022
	(0.002)	(0.001)	(0.006)	(0.008)
polviewsConservative	0.078	0.076	1.455	1.648
	(0.032)	(0.045)	(0.328)	(0.479)
polviewsLiberal	-0.104	-0.113	0.602	0.531
	(0.046)	(0.038)	(0.105)	(0.123)
(Intercept)		0.365	0.515	0.452
		(0.063)	(0.151)	(0.176)
SD (Intercept region)		0.060	1.250	1.139
SD (Observations)		0.460		

Random effects

	Region FE	Region RE	Logit	Logit + Resp
polviewsConservative	0.078	0.076	1.455	1.648
	(0.032)	(0.045)	(0.328)	(0.479)
polviewsLiberal	-0.104	-0.113	0.602	0.531
	(0.046)	(0.038)	(0.105)	(0.123)
(Intercept)		0.365	0.515	0.452
		(0.063)	(0.151)	(0.176)
SD (Intercept region)		0.060	1.250	1.139
SD (Observations)		0.460		
SD (Intercept id)				3.159
Num.Obs.	855	855	855	855
ICC		0.0	0.0	0.3

View random intercepts

The random component of the model can be extracted using the `ranef` function. This shows the point estimates for the region level deviations from the population intercept.

```
ranef(re.r)
```

```
## $region
##    (Intercept)
## 1 -0.074394507
## 2  0.016011669
## 3  0.048694832
## 4 -0.009759633
## 5  0.050229324
## 6 -0.018106992
## 7  0.035204832
## 8 -0.058255524
## 9  0.010375999
##
## with conditional variances for "region"
```

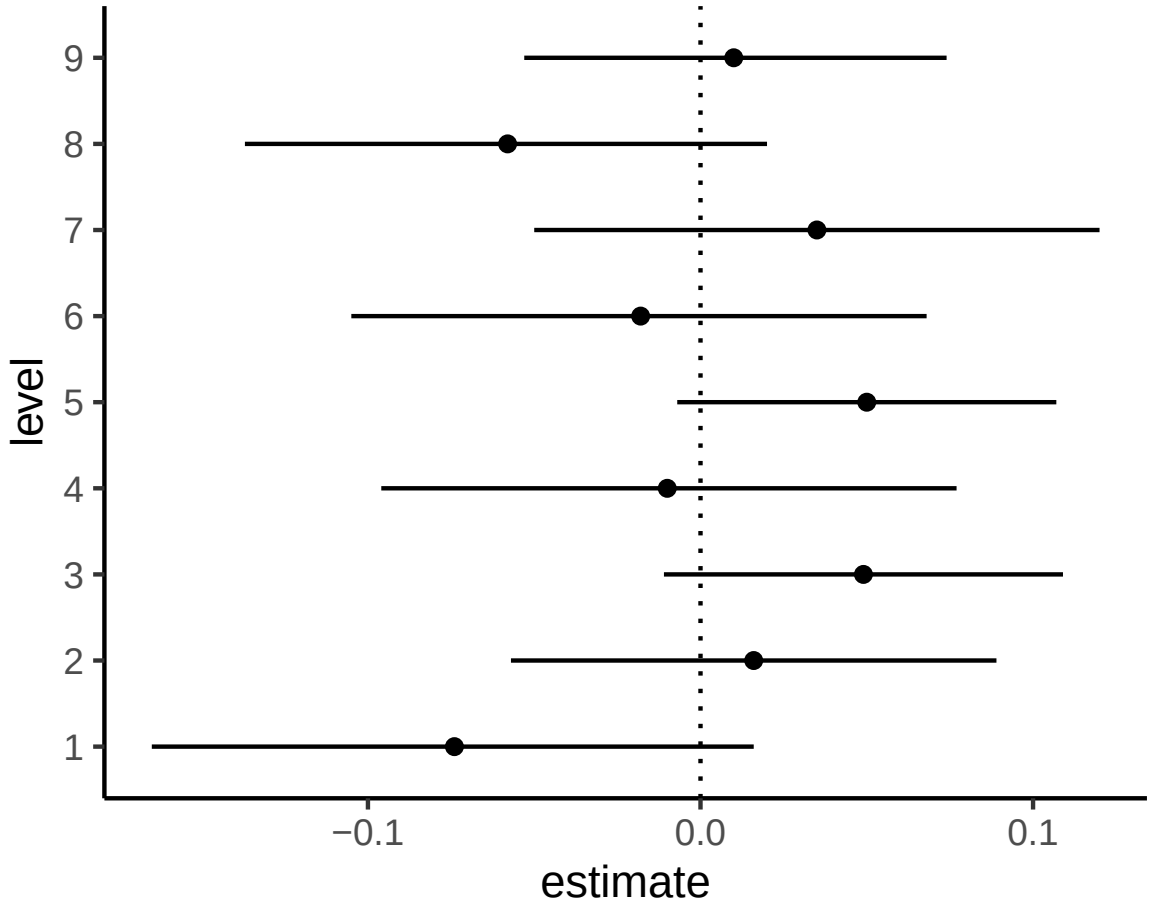
Plot random intercepts

We can get more information by using the `broom.mixed` package:

```
library(broom.mixed)
reffs <- broom.mixed::tidy(re.r, effects = "ran_vals", conf.int = TRUE) %>%
    mutate(across(where(is.numeric), round, 3)) %>%
    select(level, term, estimate, conf.low, conf.high)
reffs %>%
    head(5) %>% kable()
```

level	term	estimate	conf.low	conf.high
1	(Intercept)	-0.074	-0.165	0.016
2	(Intercept)	0.016	-0.057	0.089
3	(Intercept)	0.049	-0.011	0.109
4	(Intercept)	-0.010	-0.096	0.077
5	(Intercept)	0.050	-0.007	0.107

Plot random intercepts



Random effects

Random coefficients

- In addition to random intercepts, we can also allow the slopes to vary by group.
- For example, we might want to see whether the effect of sex on attitudes varies across regions.
- Such a model includes the population coefficient, β_{sex} and a group-level deviation $\gamma_{j,sex}$.

Random effects

Estimating random coefficient models

We can easily modify the formula to include random slopes. In this case, we allow the slope of sex to vary according to the region. The control argument is included due to estimation issues.[5]

```
rc.logit <- glmer(natcrime ~ sex + race + age + polviews +
                        (1 + sex|region),
        data = gss.new, family = binomial,
        control = glmerControl(optimizer="bobyqa", optCtrl=list(maxfun=
```

[5] Warnings suggest potential problems with the model fit that require more detailed exploration.

Random effects

	Region RE	Region RE & Race RC
(Intercept)	0.515*	0.554*
	(0.151)	(0.164)
sex2	1.928***	1.866*
	(0.297)	(0.455)
SD (Intercept region)	1.250	1.198
SD (sex2 region)		1.674
Cor (Intercept~sex2 region)		0.509
Num.Obs.	855	855
ICC	0.0	0.0

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

Random effects

```
sex.slopes <- broom.mixed::tidy(rc.logit, effects = "ran_vals", conf.in
    mutate(across(where(is.numeric), round, 3)) %>%
    filter(term == "sex2") %>%
    select(estimate, conf.low, conf.high)
sex.slopes %>% head(5) %>% kable()
```

estimate	conf.low	conf.high
-0.728	-1.435	-0.020
0.225	-0.399	0.850
0.415	-0.152	0.983
0.465	-0.311	1.240
0.195	-0.302	0.693

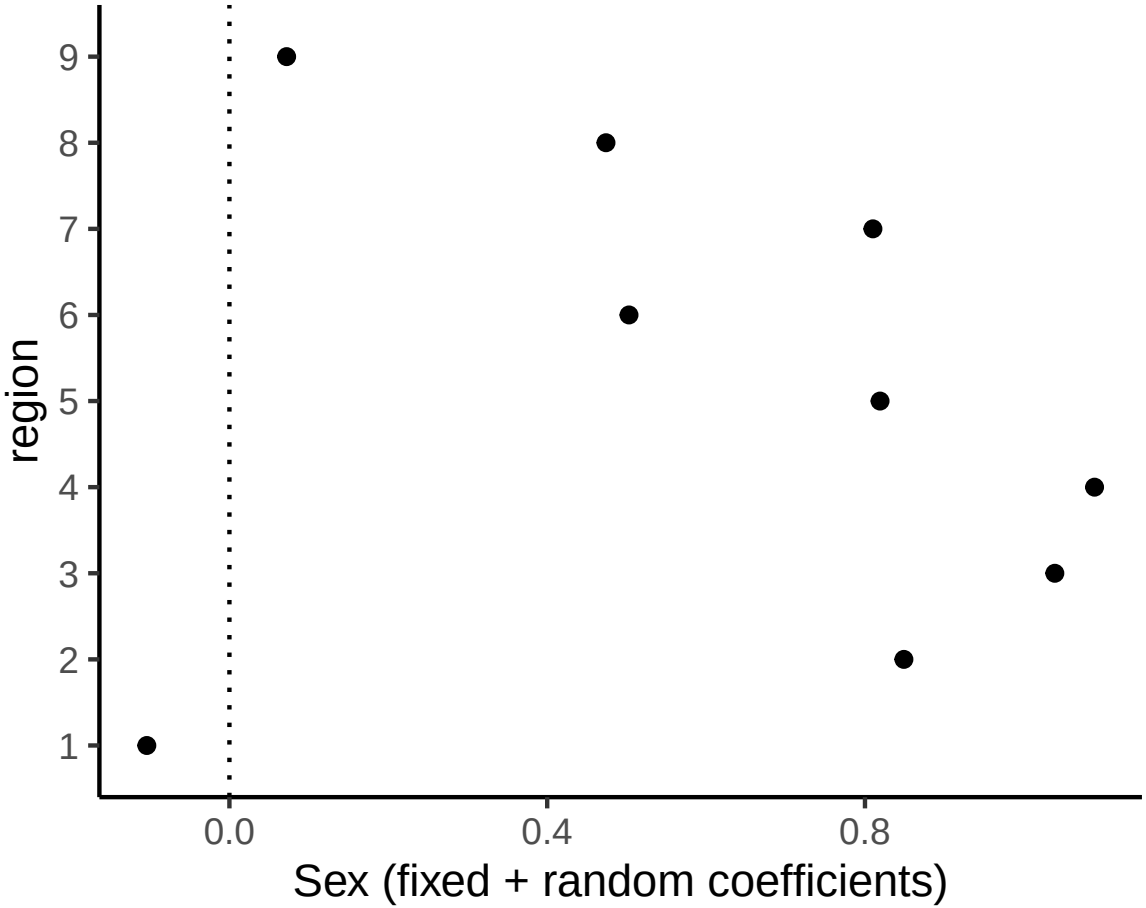
Random effects

Extracting random slopes

```
fixed.coef <- broom.mixed::tidy(rc.logit, effects = "fixed", conf.int =
    mutate(across(where(is.numeric), round, 3)) %>%
    filter(term == "sex2") %>%
    select(estimate)
est <- sex.slopes %>% mutate(sex_region = estimate + fixed.coef$estimat
```

Random effects

Extracting random slopes[6]



[6] Confidence intervals must also be calculated

Random effects

Model selection

- There is some debate over how to decide between fixed and random effects specifications.
- One convention is to use a test to identify whether random effects should be included over fixed effects using a Hausmann test, but this has been questioned (Bell and Jones 2019).
- Random effects are a more flexible approach to capture complex structures (McElreath 2020), but are not as parsimonious as fixed effects specifications.

Random effects

Advanced multilevel modeling

- Cross-level interactions can reveal relationships between different levels
  - e.g. In a model to predict child's test scores, one could interact child-level and school-level variables
- The "within-between" decomposition approach allows effects to be disentangled within and between units (see Bell and Jones 2019)
- Bayesian hierarchical modeling offers a more stable approach to complex models than MLE
  - `brms` uses the same syntax as `lme4` for model specification

Space, time and social structure

Autocorrelation

- **Autocorrelation** implies that something is correlated with itself
- Violation of IID assumption
- Unlikely to be an issue when using randomly sampled cross-sectional data, but a problem in many applied settings

Space, time and social structure

Types of autocorrelation

- Temporal autocorrelation is the most typical case, where measurements are correlated with time
  - e.g. Given quarterly GDP, we expect high correlation between GDP_t and GDP_{t-1}

Space, time and social structure

Types of autocorrelation

- Spatial autocorrelation implies that measurements are correlated with spatial proximity
  - e.g. County-level GDP more similar between proximate counties than distant ones.

Space, time and social structure

Types of autocorrelation

- Network autocorrelation implies that measurements are correlated with network position
  - This is typically a problem if we want to sample measurements from individuals who have some relationship with one another
  - e.g. Children in a classroom who are friends are more likely to have similar interests than children who are not friends ("homophily")

Space, time and social structure

Heuristics for identifying autocorrelation

- Repeated measurements
  - Temporal autocorrelation
- Spatial structure to measurements
  - Spatial autocorrelation
- Non-random or network sampling
  - Network autocorrelation

Space, time and social structure

Solutions

- Standard error corrections
  - Appropriate error structures
- Fixed and random effects
  - Directly model data structure
- Data processing
  - e.g. De-trending and de-seasoning time series variables
- Model specification
  - e.g. Lagged variables, differences, spatial autocorrelation terms
- More advanced approaches
  - ERGM and SAOM models for networks

Space, time and social structure

Takeaways

- Standard GLMs alone are often insufficient to account for the way data are structured
- Standard error corrections are often necessary, but not a panacea
- Fixed effects and random effects models allow structure to be modeled in different ways
- More complex types of structure and dynamics should be directly modeled to avoid misleading inferences

Causal inference and regression

Potential outcomes and the fundamental problem of causal inference

- D_i is a binary variable denoting whether a unit is treated.
- For each unit, we have two **potential outcomes**:
  - Outcome if $D_i = 1$ (treated): Y_i^1
  - Outcome if $D_i = 0$ (untreated): Y_i^0
- We want to infer the difference in potential outcomes across treatments for a given unit:
  - $Y_i^1 D_i + Y_i^0 (1 - D_i)$
- But one of the outcomes is never observed. This is known as a **counterfactual**.

Causal inference and regression

Introduction to causal inference

- In an experiment, we can randomly assign subjects to treatment and control conditions and compare the outcomes across subjects.
- Assuming a binary treatment, D we could estimate the following regression:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 D_i + u_i$$

- Here $\hat{\beta}_1$ will provide an estimate of the average treatment effect.

Causal inference and regression

Observational data

- In many settings of social scientific interest we do not have well controlled experiments. Nonetheless, we might want to consider some variables as treatments. e.g. What is effect of college degree on earnings?
- Assignment to treatment is not controlled by researcher or randomized.
- This raises the possibility of **selection bias**, as subjects may select into treatment.

Causal inference and regression

Observational data

- There are several approaches to making causal inference using observational data including
  - (Propensity score) matching and weighting
  - Instrumental variables
  - Regression discontinuity
  - Difference-in-difference

Causal inference and regression

Matching

- Intuition: Find treated and untreated units with similar covariates then compare outcomes.
- Exact matching
  - Ideal case, but limited value in practice as requires extreme subsampling
- Partial or fuzzy matching
  - Compromise, but better sample properties
- Propensity score matching/weighting
  - Estimate a model $\hat{D}_i = \hat{\beta} X_i + \hat{u}_i$
  - Use $\hat{D}_i$ to match units or weight regression equation[7]

[7] The use of PSM has been strongly criticized, see King, Gary, and Richard Nielsen. 2019. "Why Propensity Scores Should Not Be Used for Matching." *Political Analysis* 27(4):435–54. doi: 10.1017/pan.2019.11

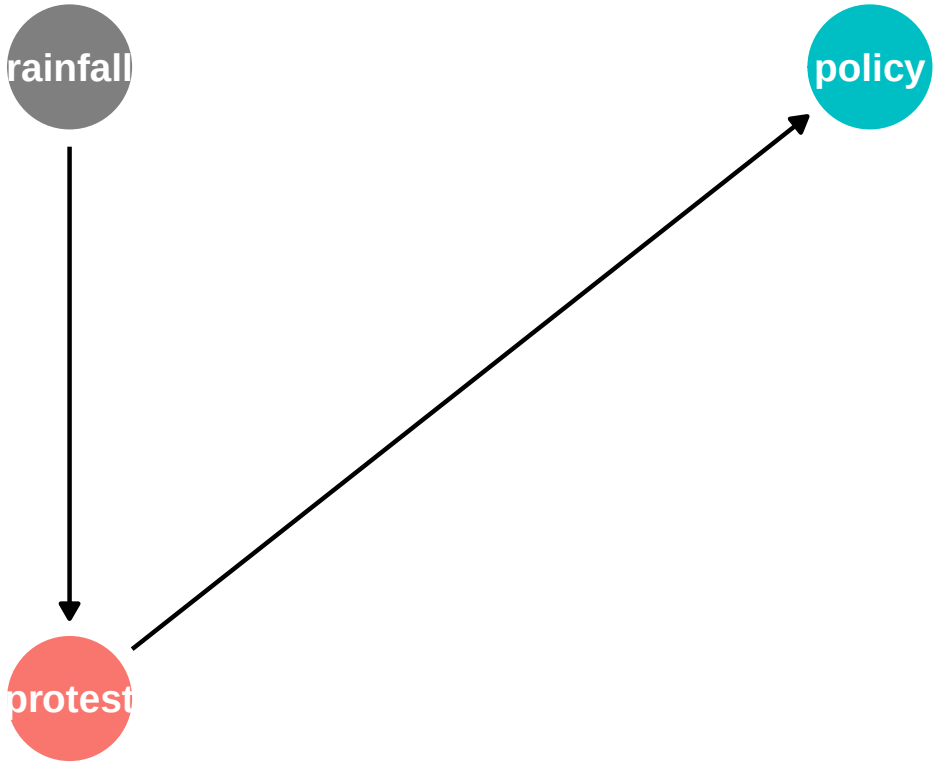
Causal inference and regression

Instrumental variables

- Intuition: Identify an exogenous regressor that can be used to explain random variation in a treatment
- Example: Rainfall and protest
  - We want to infer effect of protest (treatment) on policy change (outcome)
  - But protest is not randomly assigned
  - Rainfall is an **instrument** insofar as it effects protest and indirectly effects policy change through its effect on protest

Causal inference and regression

Instrumental variables



Causal inference and regression

Instrumental variables and two-stage least squares

- We can estimate this relationship using two-stage least squares (2SLS)
- Where Y is the outcome, D is the treatment, and Z is the instrument:
  - First stage: $\hat{D}_i = \hat{\gamma}_0 + \hat{\gamma}_1 Z_i + u_i$
  - Second stage: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1(\hat{D}_i) + u_i$
- Note: Additional controls can be included in both stages.

Causal inference and regression

IV assumptions and requirements

- Assumptions
  - Z has a causal effect on D (relevance)
  - Z effects Y only through D (exclusion restriction)
  - Z and Y do not share common causes (independence)
- Requirements
  - F-statistic in first-stage should be greater than 10, otherwise considered a **weak instrument**
  - A large econometric literature on diagnostics and assumptions

Summary

- IID assumptions often violated and data structures must be accounted for
- A variety of statistical techniques can be used to explicitly model these structures
- Causal inference is difficult in observational settings due to selection bias
- Various regression-based techniques can be used to infer causality