SOC542 Statistical Methods in Sociology II

Fixed effects, random effects, and autocorrelation

Thomas Davidson

Rutgers University

April 21, 2025

Course updates

- Project updates due Friday 4/25 at 5pm via email
  - Descriptive statistics: One or more tables or figures
  - Regression analyses: One or more regression tables showing
    - Bivariate results
    - Multivariate results
  - At least one figure showing estimates from regression (e.g. coefficients, predictions, marginal effects)
  - Draft methodology and results sections

Course updates

- Presentations on 5/5
  - 10 minutes to present project
    - Introduction
    - Data
    - Methodology
    - Main results
    - Robustness checks
    - Conclusions
  - 5 minutes for Q&A

Plan

- Violations of regression assumptions
- Robust and clustered standard errors
- Fixed effects
- Random effects
- Autocorrelation: space, time, and structure

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 (IID)
  - 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
- Type I errors (false positives) more likely
- 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]

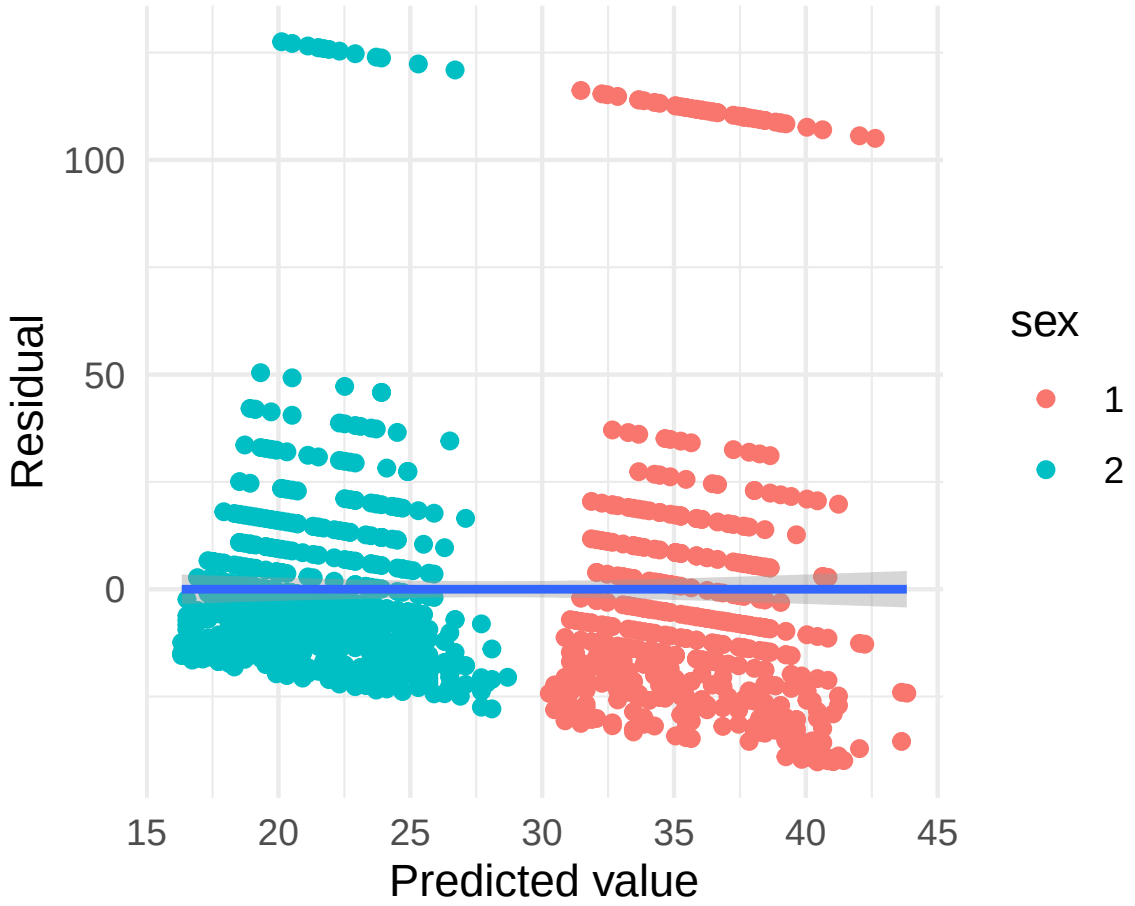
[1] See King and Roberts 2015 for further technical discussion.

Robust and clustered standard errors

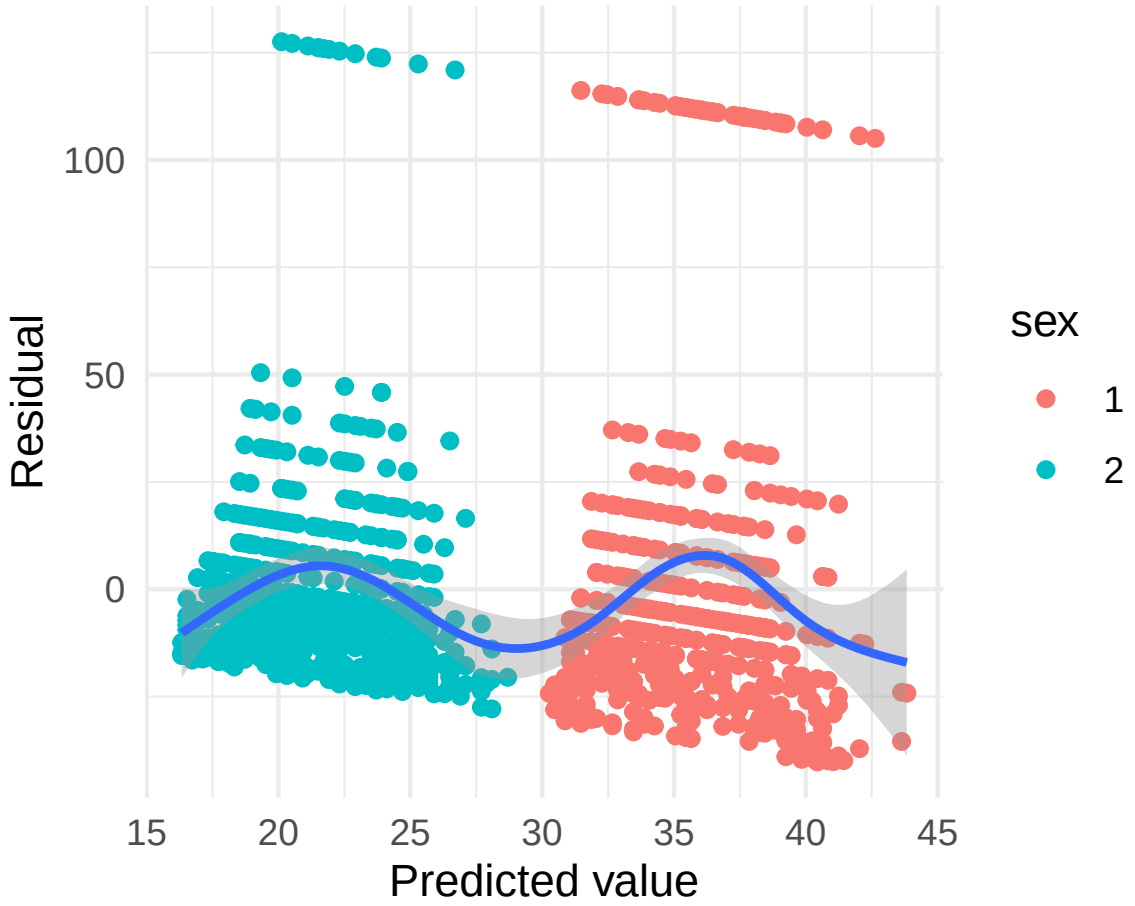
Estimating a simple model: Income as a function of age and sex (GSS 2020)

```
m <- lm(I(realrinc/1000) ~ 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)	26.285***	26.285***	26.285***
	(3.458)	(3.063)	(1.929)
age	0.199**	0.199**	0.199***
	(0.066)	(0.062)	(0.040)
sex2	-13.941***	-13.941***	-13.941***
	(1.915)	(1.961)	(0.068)
Num.Obs.	1077	1077	1077

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

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.

Fixed effects

- A fixed effects model can be written like a standard regression model.

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

- γ is a vector of fixed effect coefficients, one dummy variable for each group.
- The γ_j term absorbs unexplained variance in group j.
- It is common to drop the global intercept.

Fixed effects

Pooling

- **Pooling** refers to how observations are pooled together to estimate averages.
- Considering data with a grouped structure, like students sampled from schools
  - 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

Panel data

- GSS panel
  - Sample of 2016 and 2018 respondents were re-interviewed in 2020 (online)
  - Each row is one person-year

Data and Methodology

- Dependent variable
  - `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)
- Independent variable
  - Ideology
- Controls
  - Sex, age, race
- LPM with fixed effects
  - Survey year
  - Region

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 = gss.new)
fe.y <- feols(natcrime ~ sex + race + age + polviews | year,
              data = gss.new)
fe.ry <- feols(natcrime ~ sex + race + age + polviews | region + year,
               data = gss.new)
fe.ryi <- feols(natcrime ~ sex + race + age + polviews | region + year + id,
               data = gss.new)
```

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

Fixed effects

	Pooled	Region FE	Year FE	R-Y	R-Y-Indiv.
(Intercept)	0.379				
	(0.060)				
polviewsConservative	0.077	0.078	0.076	0.077	0.090
	(0.045)	(0.032)	(0.018)	(0.032)	(0.101)
polviewsLiberal	-0.119	-0.104	-0.119	-0.104	-0.018
	(0.038)	(0.046)	(0.016)	(0.046)	(0.105)
Num.Obs.	855	855	855	855	855
R2	0.070	0.090	0.071	0.092	0.682
R2 Adj.	0.063	0.075	0.062	0.074	0.315

Fixed effects

Interpretation

- The fixed effects account 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

Incorporating clustered standard errors

We can modify the arguments of `feols` to modify the way standard errors are calculated. In this case, they are being clustered by respondent ID.

Fixed effects

	R-Y	R-Y FE (Clustered)
polviewsConservative	0.077	0.077
	(0.032)	(0.042)
polviewsLiberal	-0.104	-0.104
	(0.046)	(0.043)
Num.Obs.	855	855
R2	0.092	0.092
R2 Adj.	0.074	0.074

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

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

Where we assume that the error term has a normal distribution

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

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

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

Random effects

Partial pooling

- 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.

Random effects

Partial pooling and shrinkage

- 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 researchers 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.

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

Random effects

	Region FE	Region RE	Logit Region RE	Logit-R-Indiv.
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

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"
```

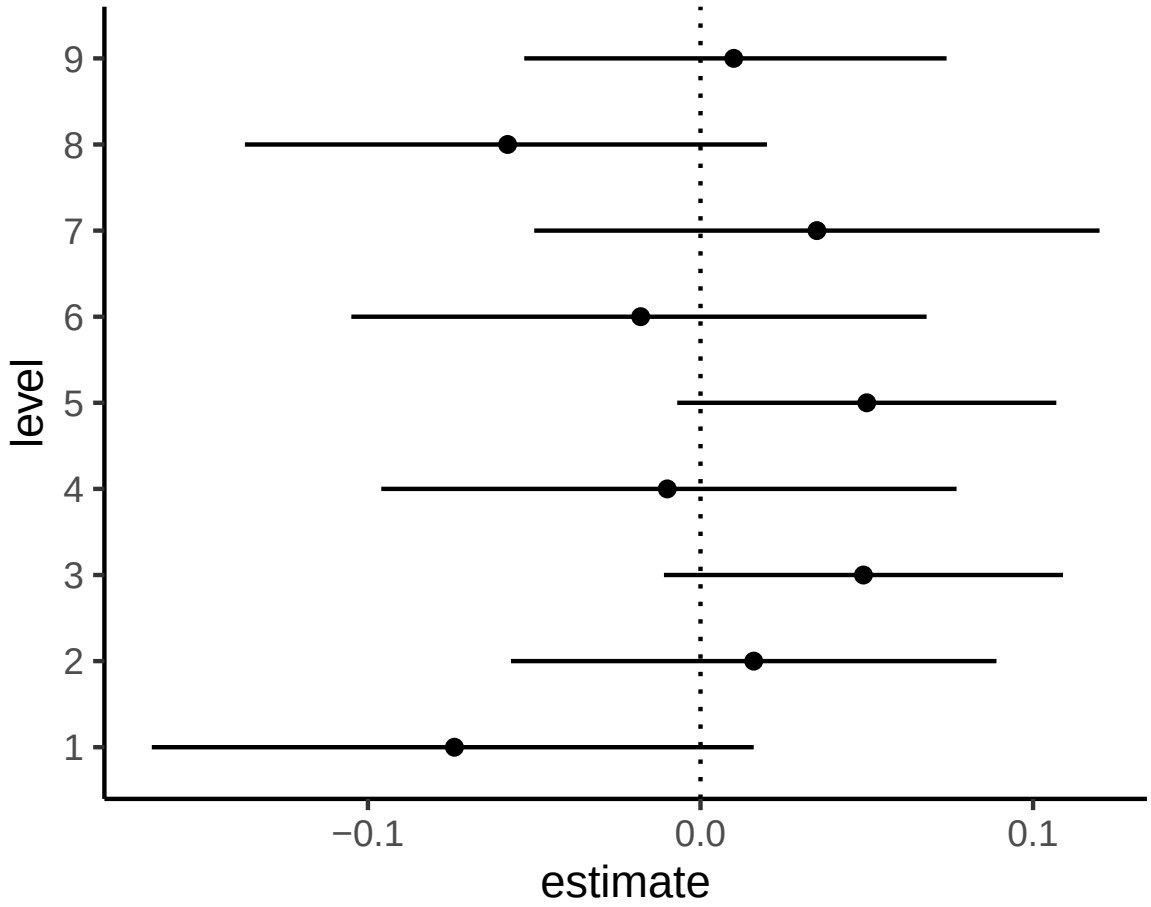
Plot random intercepts

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, does 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. The control argument is included due to estimation issues.[4]

```
rc.logit <- glmer(natcrime ~ sex + race + age + polviews +
                        (1 + sex|region),
        data = gss.new, family = binomial,
        control = glmerControl(optimizer="bobyqa",
                               optCtrl=list(maxfun=2e5)))
```

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

Random effects

	Region RE	Region RE Sex 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 <- tidy(rc.logit, effects = "ran_vals", conf.int = TRUE) %>%
    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

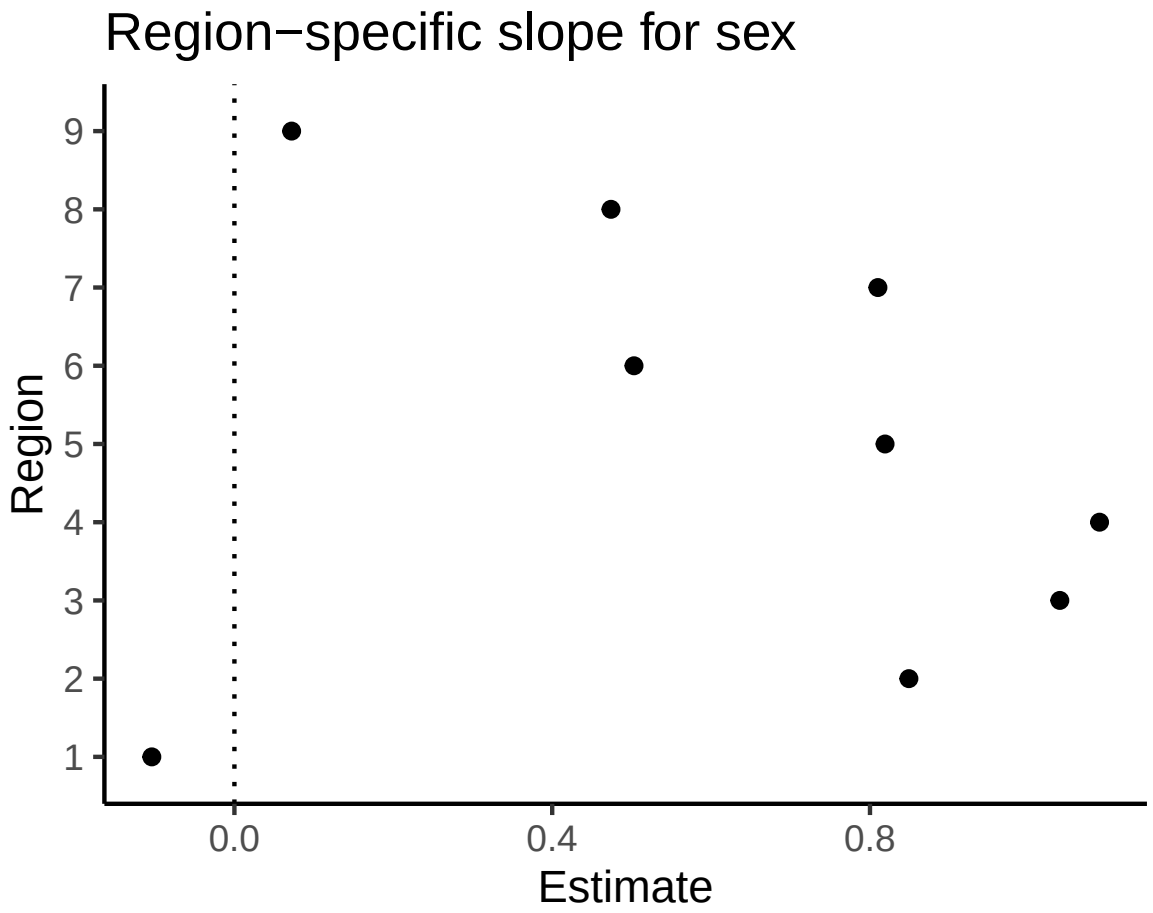
Random effects

Extracting random slopes

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

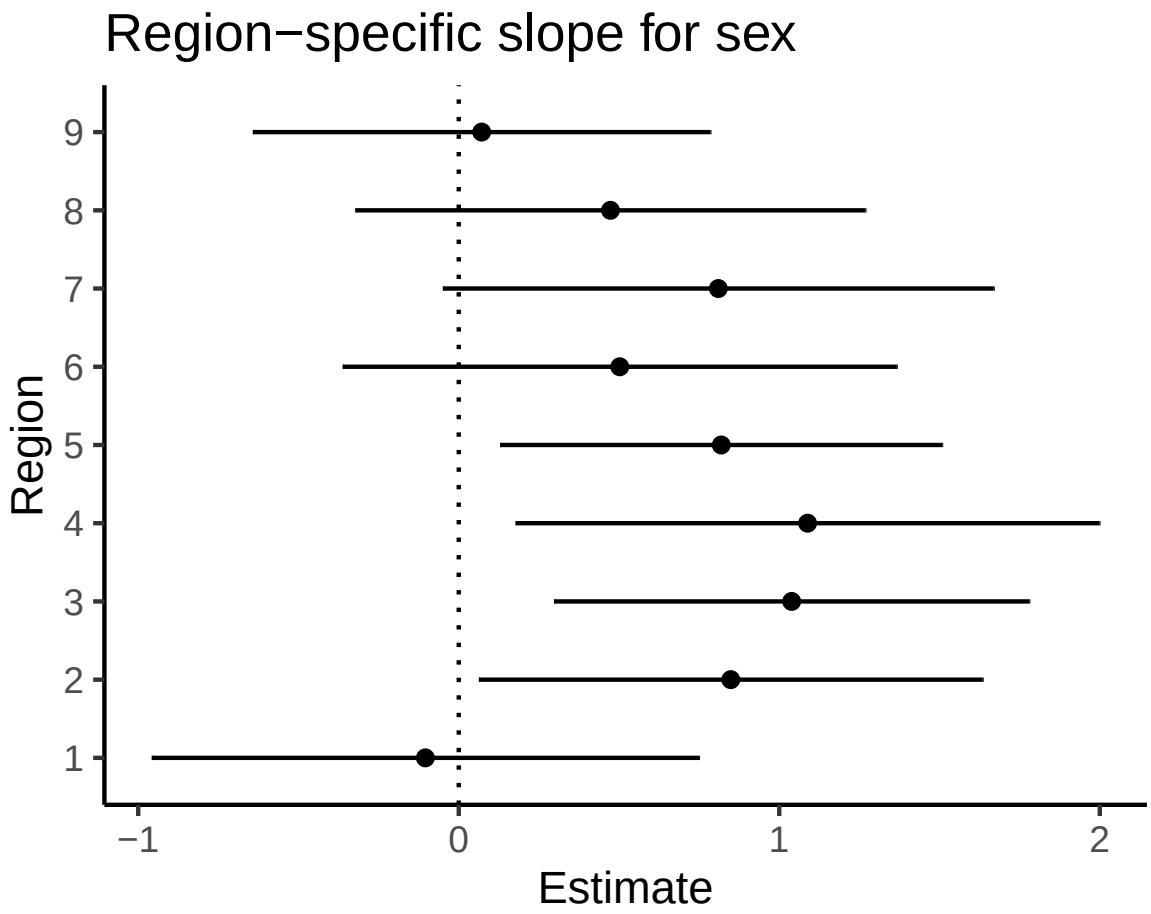
Random effects

Extracting random slopes



Random effects

Extracting random slopes with confidence intervals



Choosing between fixed and random effects

- Some sociologists argue that random effects are inappropriate due to "heroic" assumption that the random effects are uncorrelated with the predictors (Vaisey and Miles 2017).
  - Exception is where group assignment can be assumed to be as good as random.
- One convention is to use a test to identify whether random effects should be included over fixed effects using a Hausmann test, which evaluates whether covariates are correlated with the random effects, but this practice been questioned (Bell and Jones 2019).
- Transformations and Bayesian approaches can address confounding in multilevel settings (McElreath 2020, see also McElreath's 2023 Lecture 12 on YouTube)

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
  - Priors can be specified for complex correlation structures

Space, time and social structure

Autocorrelation

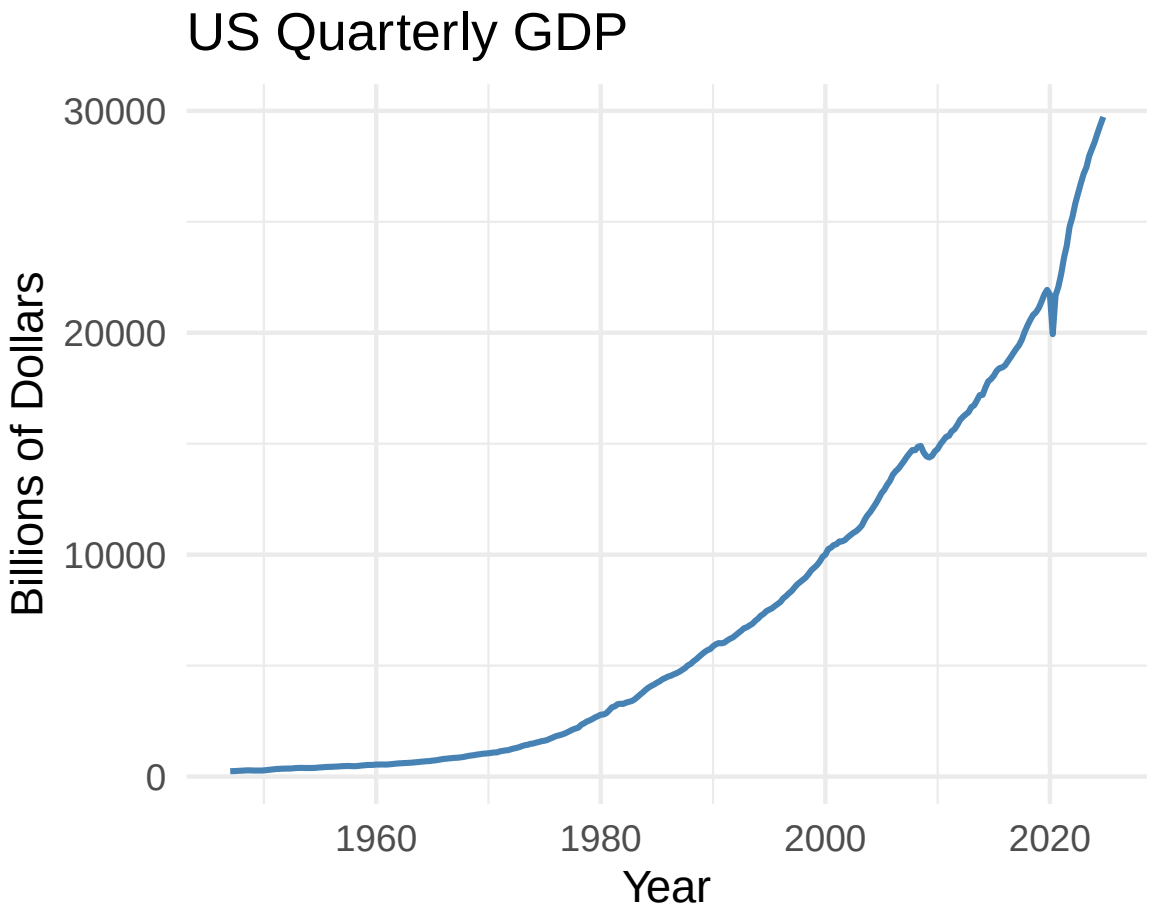
- **Autocorrelation** implies that something is correlated with itself
- Violation of IID assumption
- Unlikely to be an issue when using randomly sampled cross-sectional data, but a problem in many applied settings

Space, time and social structure

Types of autocorrelation

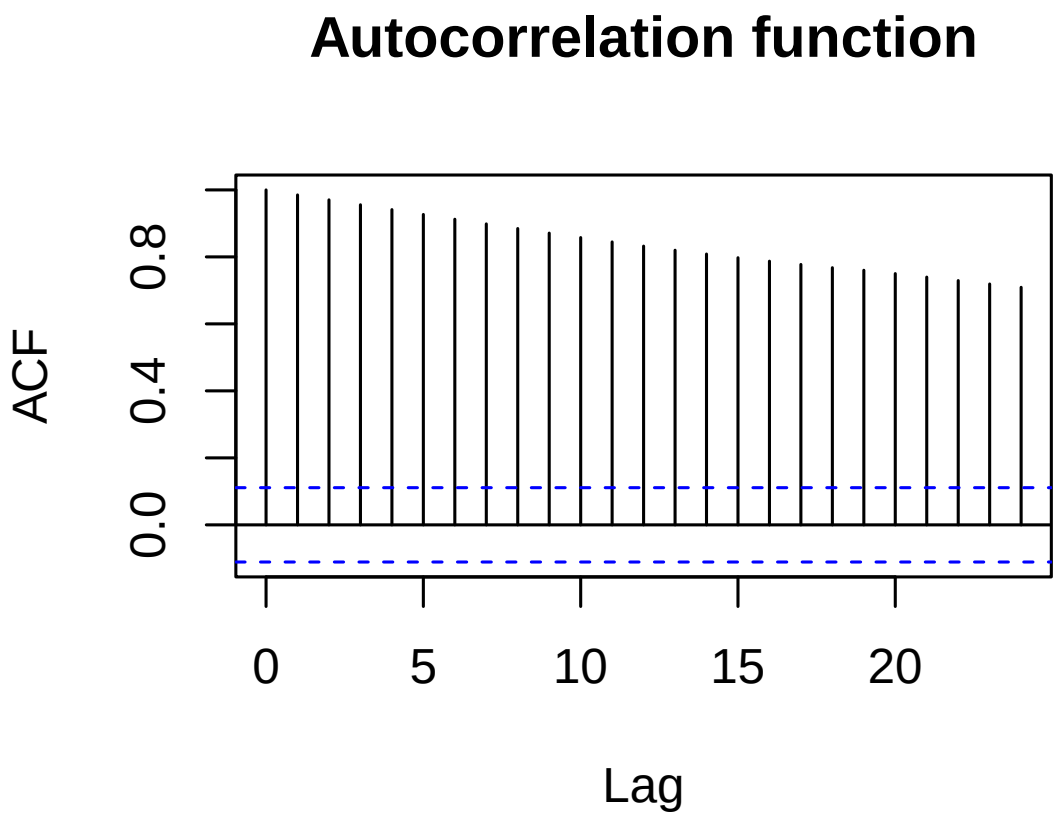
- Temporal autocorrelation is the most typical case, where measurements are correlated with time
  - e.g. Given quarterly GDP, we expect high correlation between GDP_t and GDP_{t-1}

Temporal autocorrelation



Temporal autocorrelation

```
acf(gdp_df$value, main = "Autocorrelation function")
```

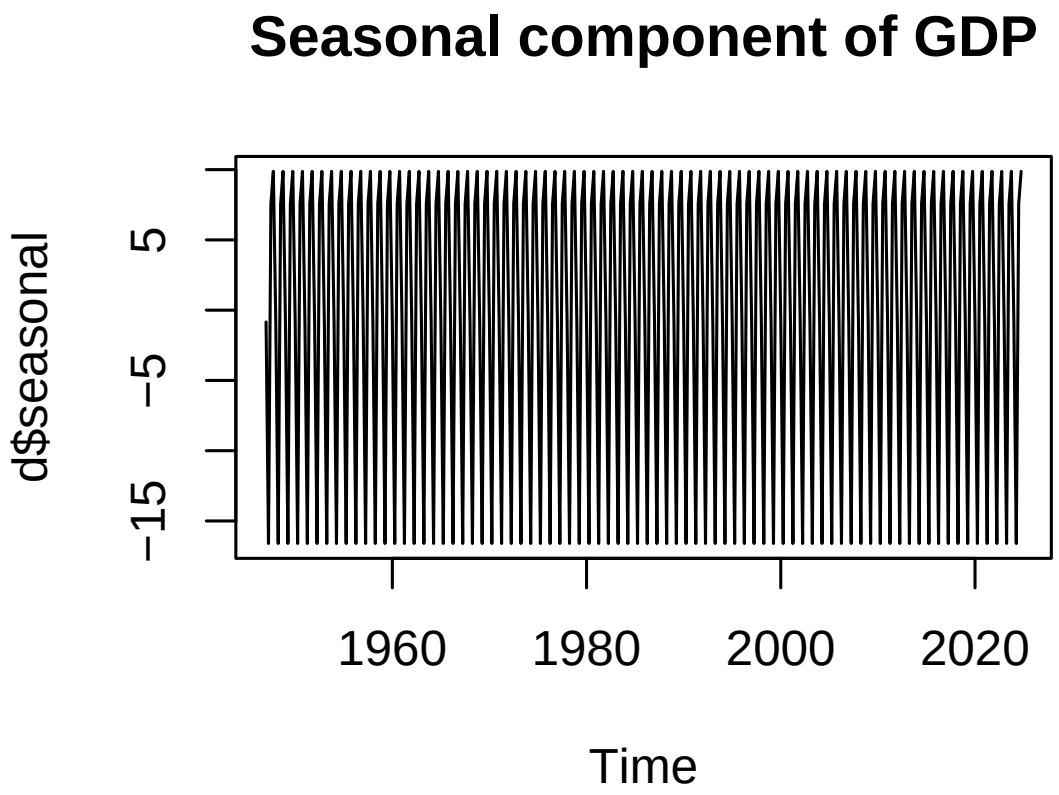


Time series decomposition

- Time series data can be decomposed into different components
  - Trend represents the long-term movement in the data
  - Seasonality
  - The "random" component is the residual variation that cannot be explained by the trend or seasonality

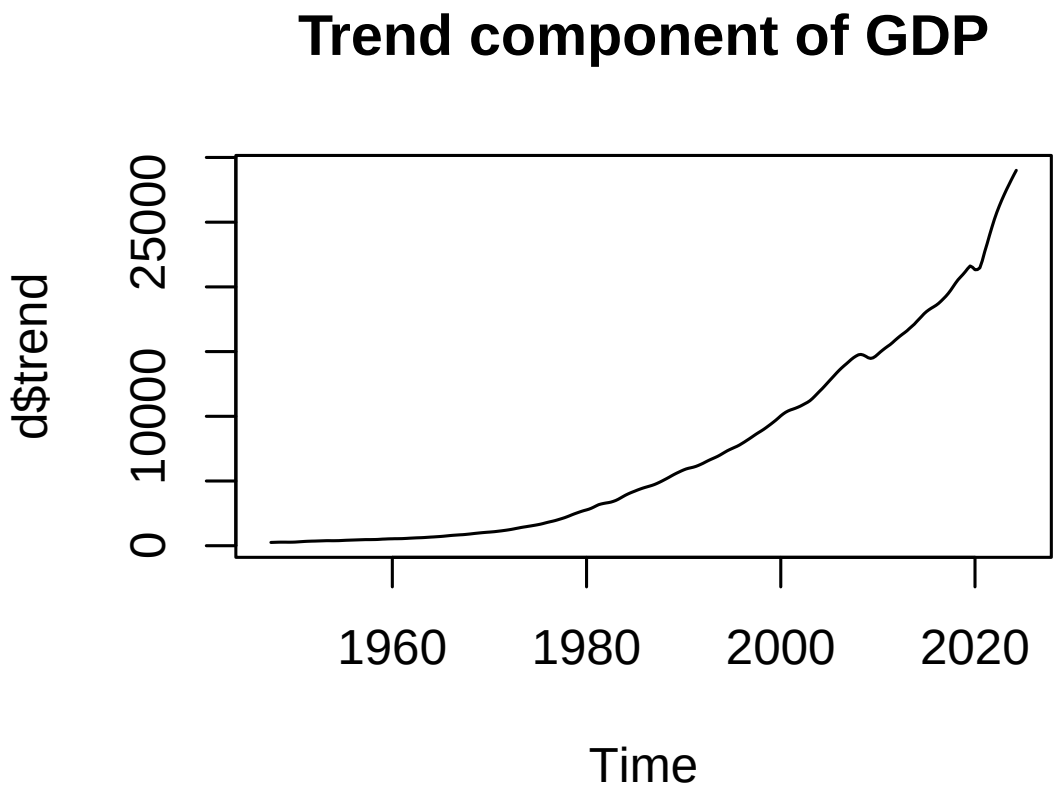
Time series decomposition

Seasonality



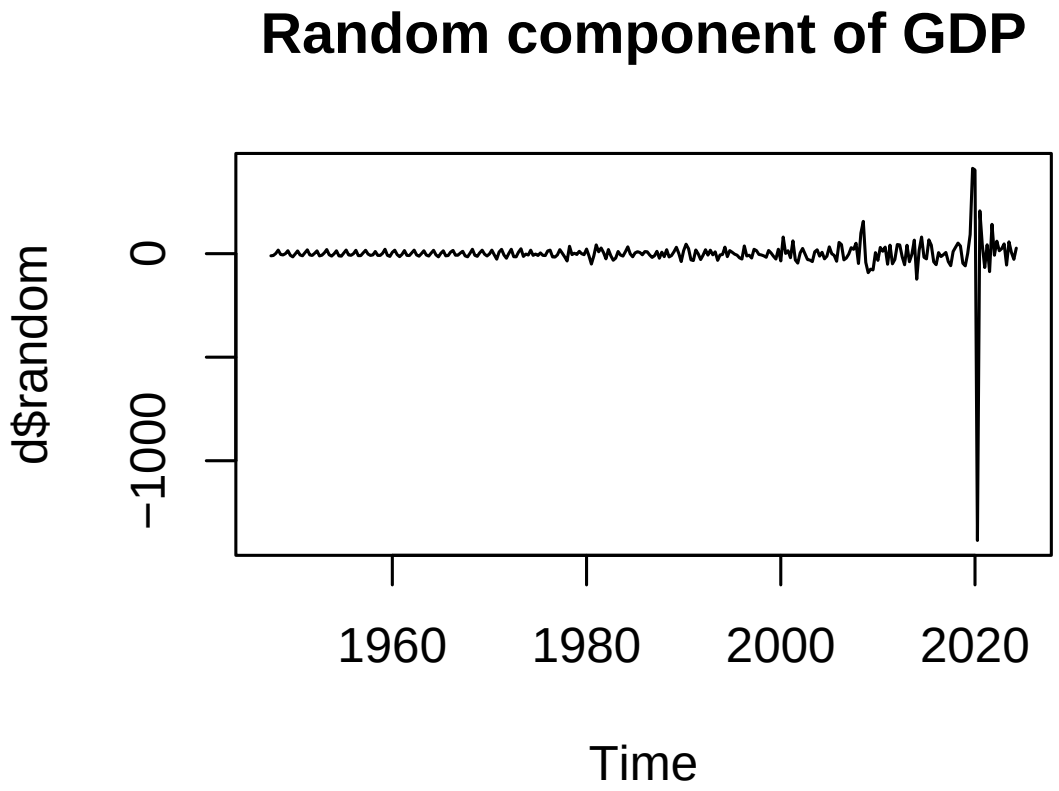
Time series decomposition

Trend



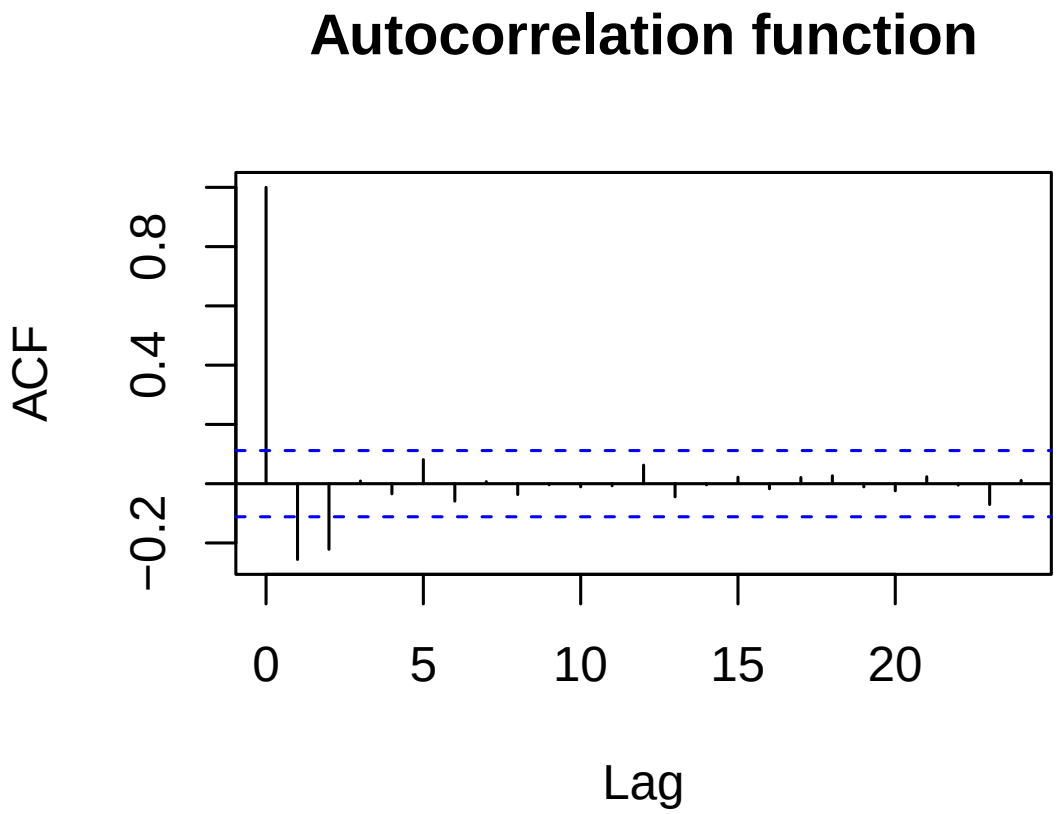
Time series decomposition

Random



Temporal autocorrelation

Random component



Addressing autocorrelation

- Solution: Model the temporal structure
- The AR(1) model:

$$y_t = \alpha + \beta y_{t-1} + \varepsilon_t$$

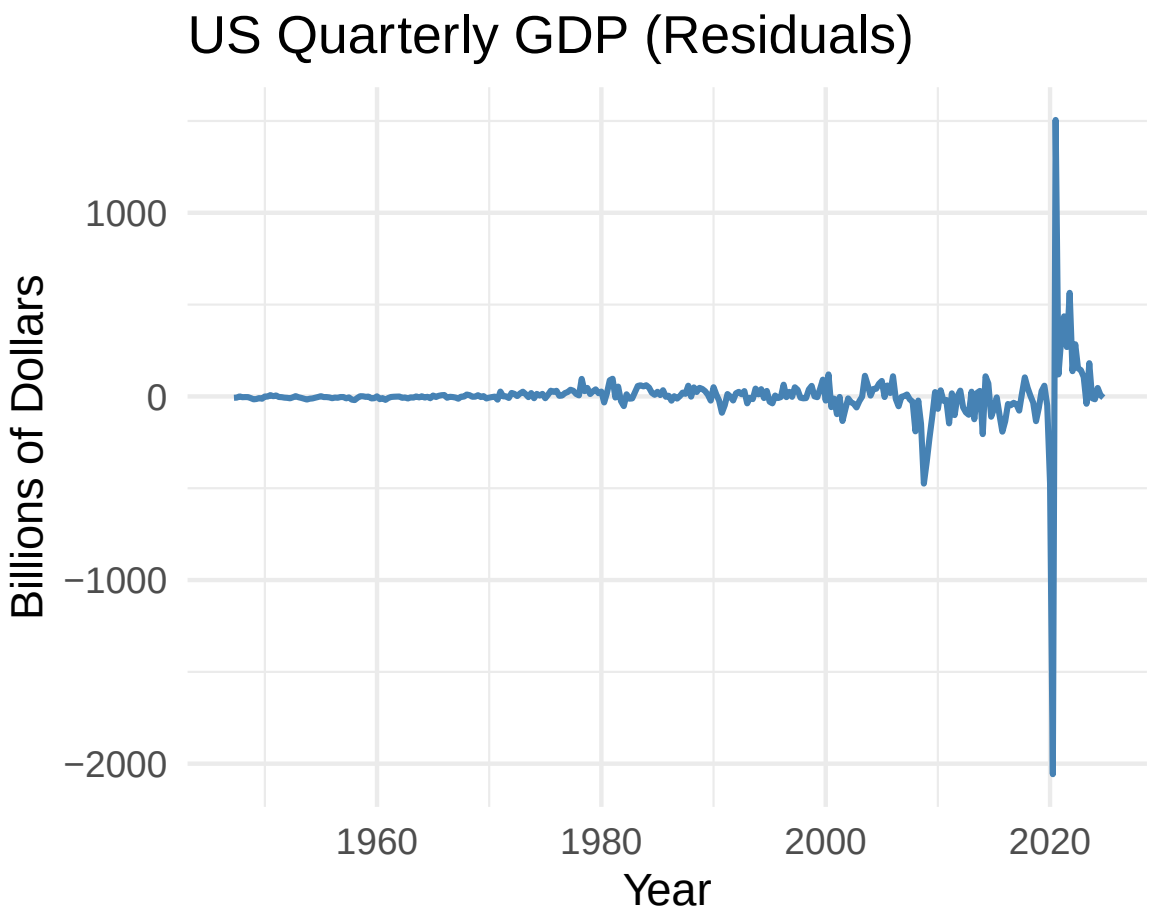
- Past values help predict the present - In this case, GDP at quarter t is explained by GDP at quarter $t-1$.

Addressing autocorrelation

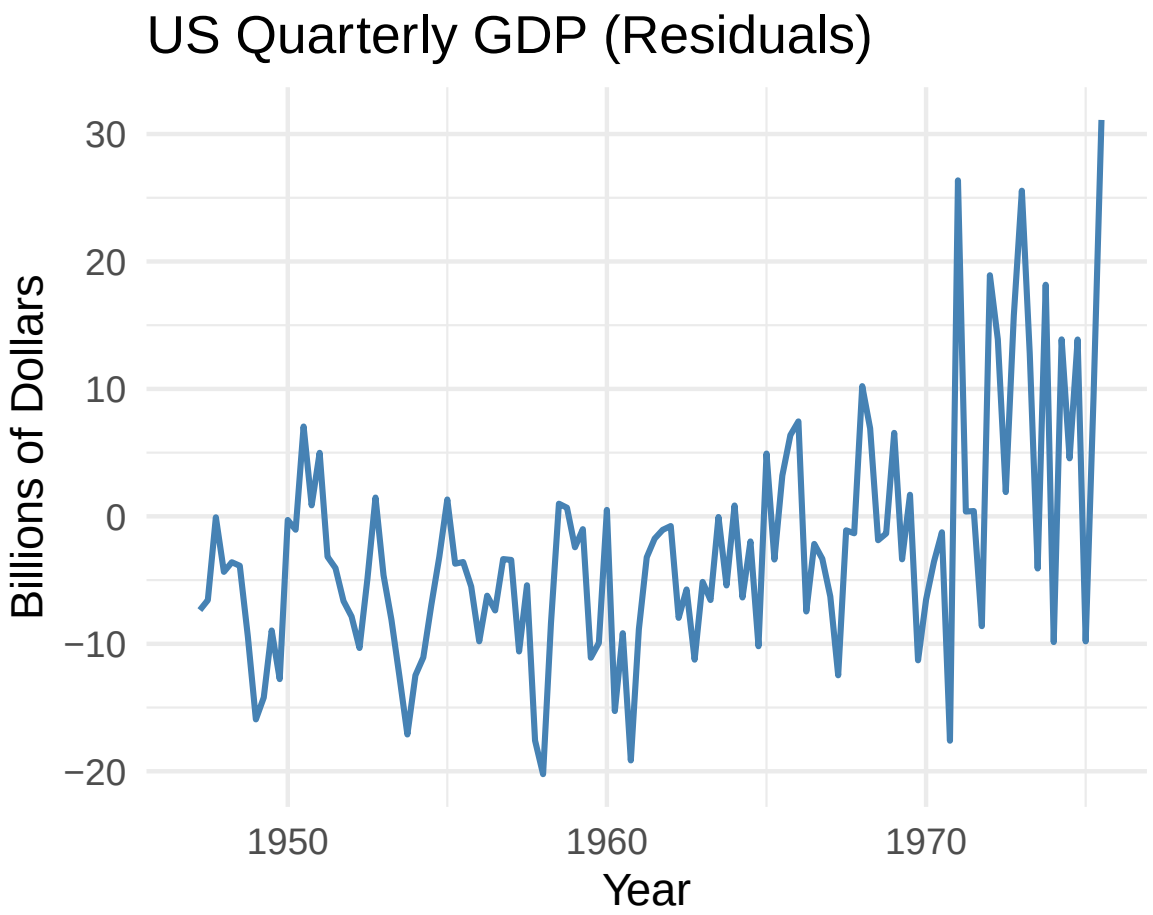
	(1)
(Intercept)	7.274
	(13.131)
lag(value)	1.012***
	(0.001)
Num.Obs.	311
R2	1.000
R2 Adj.	1.000
F	673023.000

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

Addressing autocorrelation

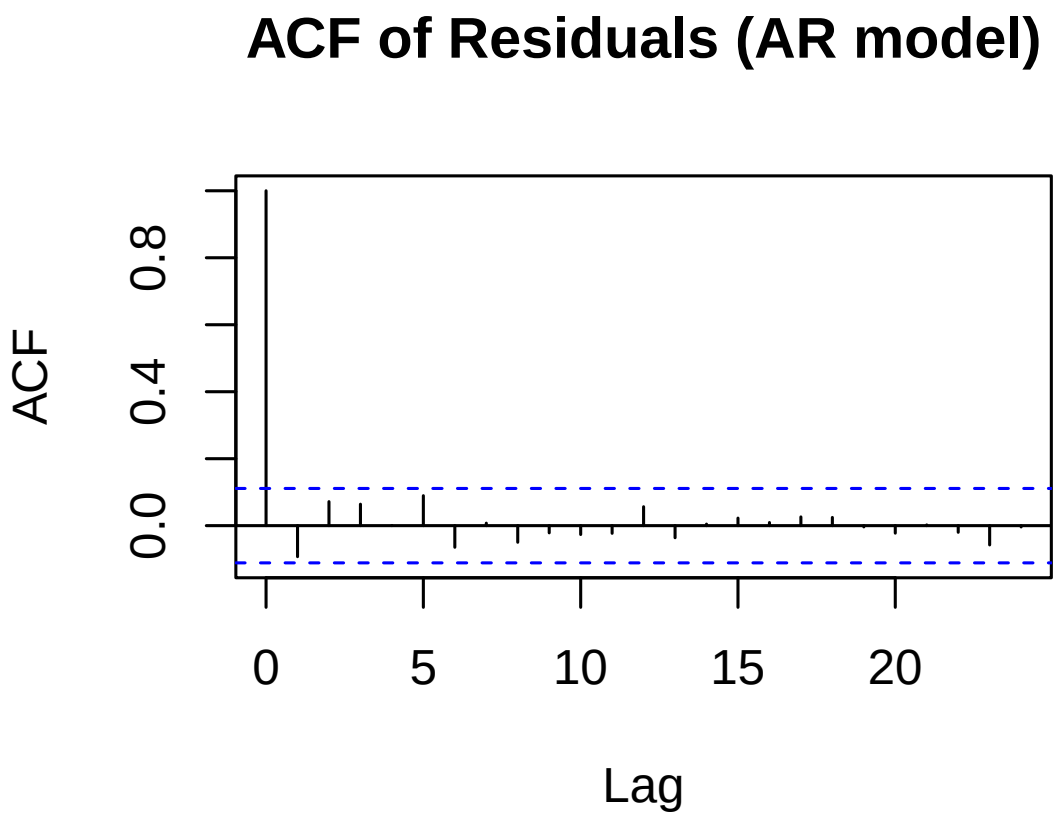


Addressing autocorrelation



Addressing autocorrelation

```
acf(residuals(ar_model), main = "ACF of Residuals (AR model)")
```



Time series analysis

- Time series analysis is a branch of statistics involved in analyzing time series
  - Simple AR(1) models can be used in some settings, but more complex models are often necessary in multivariate settings
  - Additional properties of series that must be modeled include
    - Trends and seasonality
    - Non-stationarity (changing mean and variance over time)

Testing for stationarity

```
library(tseries)
adf.test(gdp_df$value)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  gdp_df$value
## Dickey-Fuller = 2.6247, Lag order = 6, p-value = 0.99
## alternative hypothesis: stationary
```

```
adf.test(gdp_df$residuals[-1])
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  gdp_df$residuals[-1]
## Dickey-Fuller = -6.2211, Lag order = 6, p-value = 0.01
## alternative hypothesis: stationary
```

Space, time and social structure

Types of autocorrelation

- Spatial autocorrelation implies that measurements are correlated with spatial proximity
  - e.g. County-level population more similar between proximate counties than distant ones.
- Spatial regression methods provide ways to account for this when using spatial data

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")
- Network autocorrelation can be modeled using Exponential Random Graph Models (ERGM) or Stochastic Actor-Oriented Models (SAOM)

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

Summary

- IID assumptions often violated when analyzing structured data
- Fixed effects can absorb unobserved heterogeneity across units
  - No pooling
  - Perfect multicollinearity
- Random effects can be used to model more complex structures
  - Partial pooling and shrinkage
  - Random slopes
- Autocorrelation is a common problem in structured data
  - Temporal, spatial, and network autocorrelation
  - A variety of statistical techniques can be used to directly model these structures