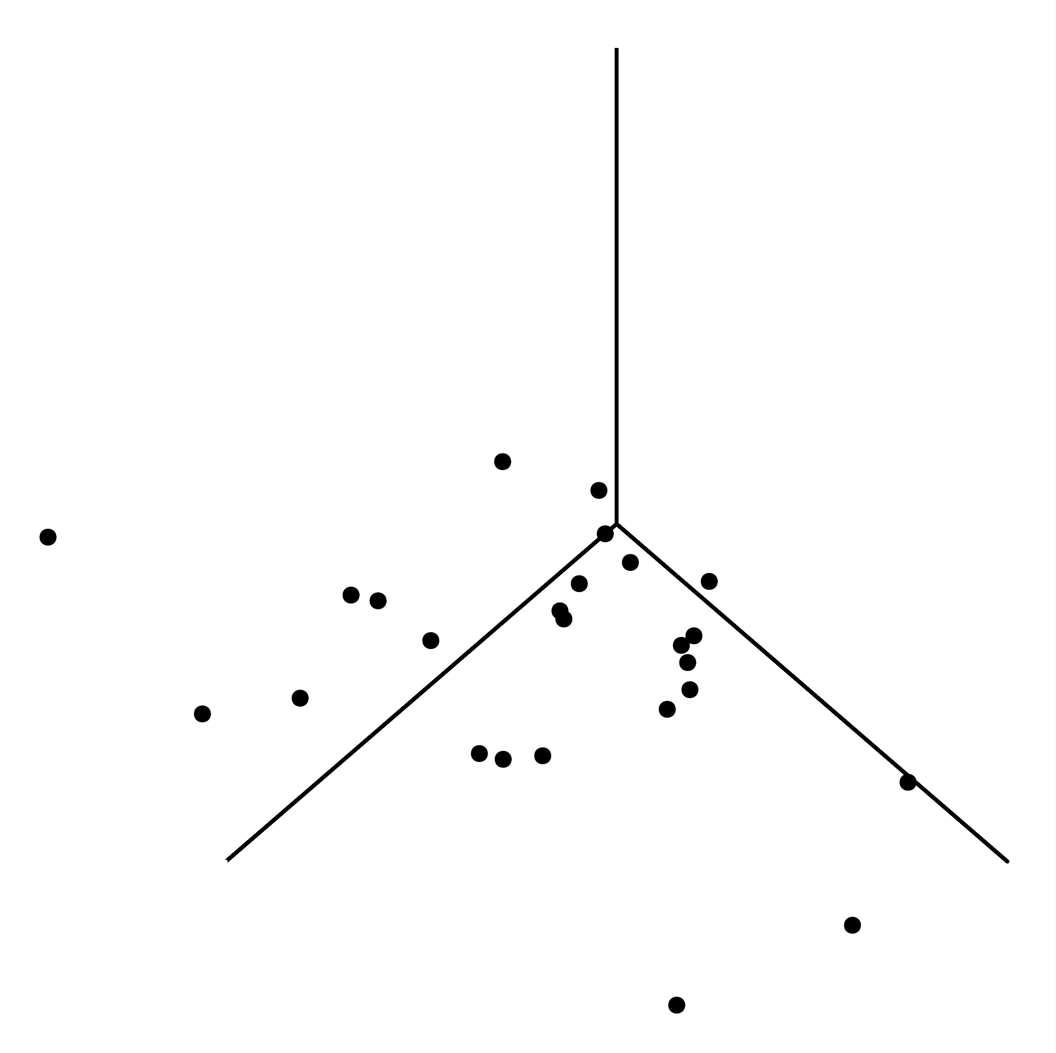
Serial Correlation

Stat 230

May 06 2022

Overview



Today:

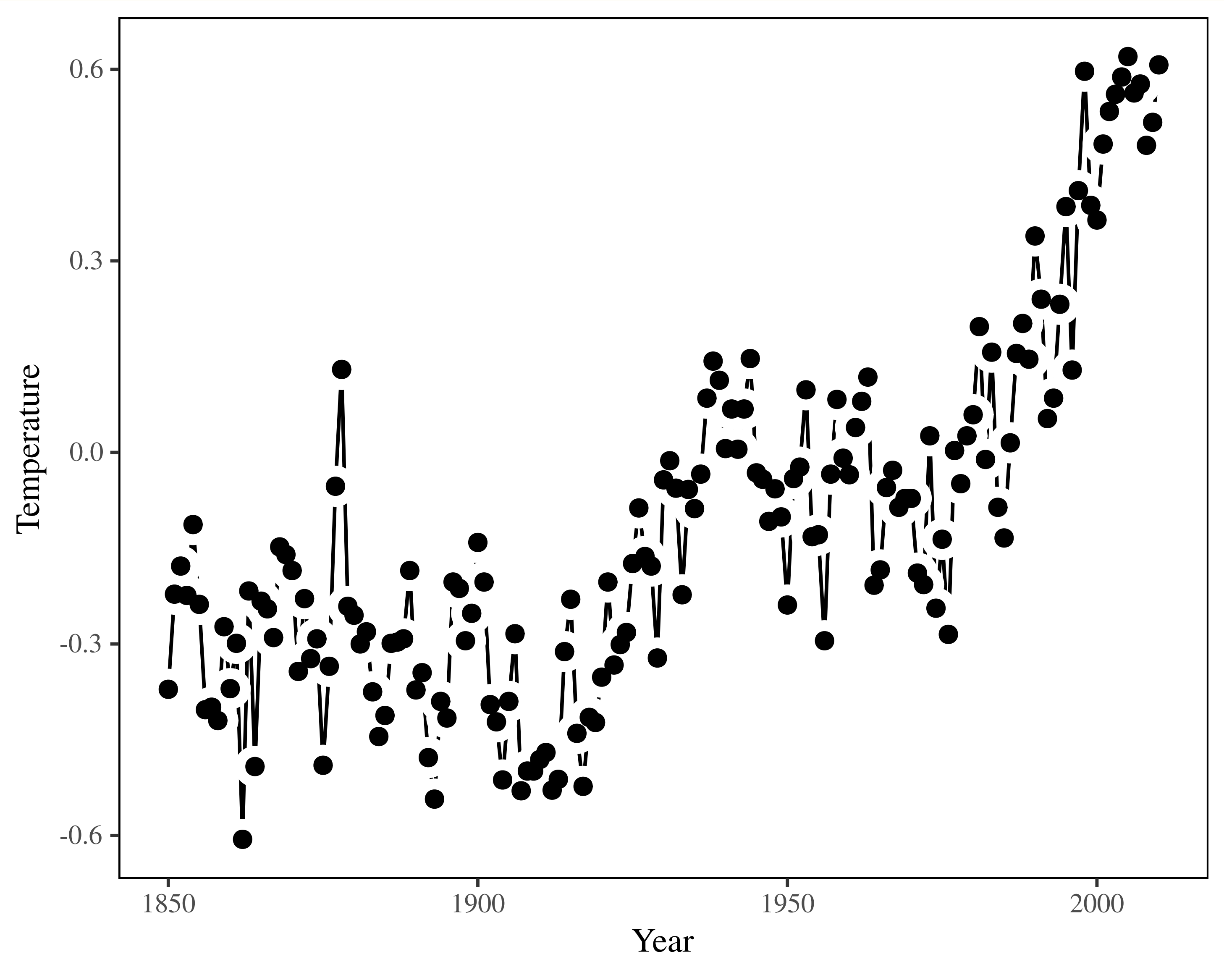
- Violation of independence
- Serial correlation
- Partial autocorrelation

Measuring Global Warming

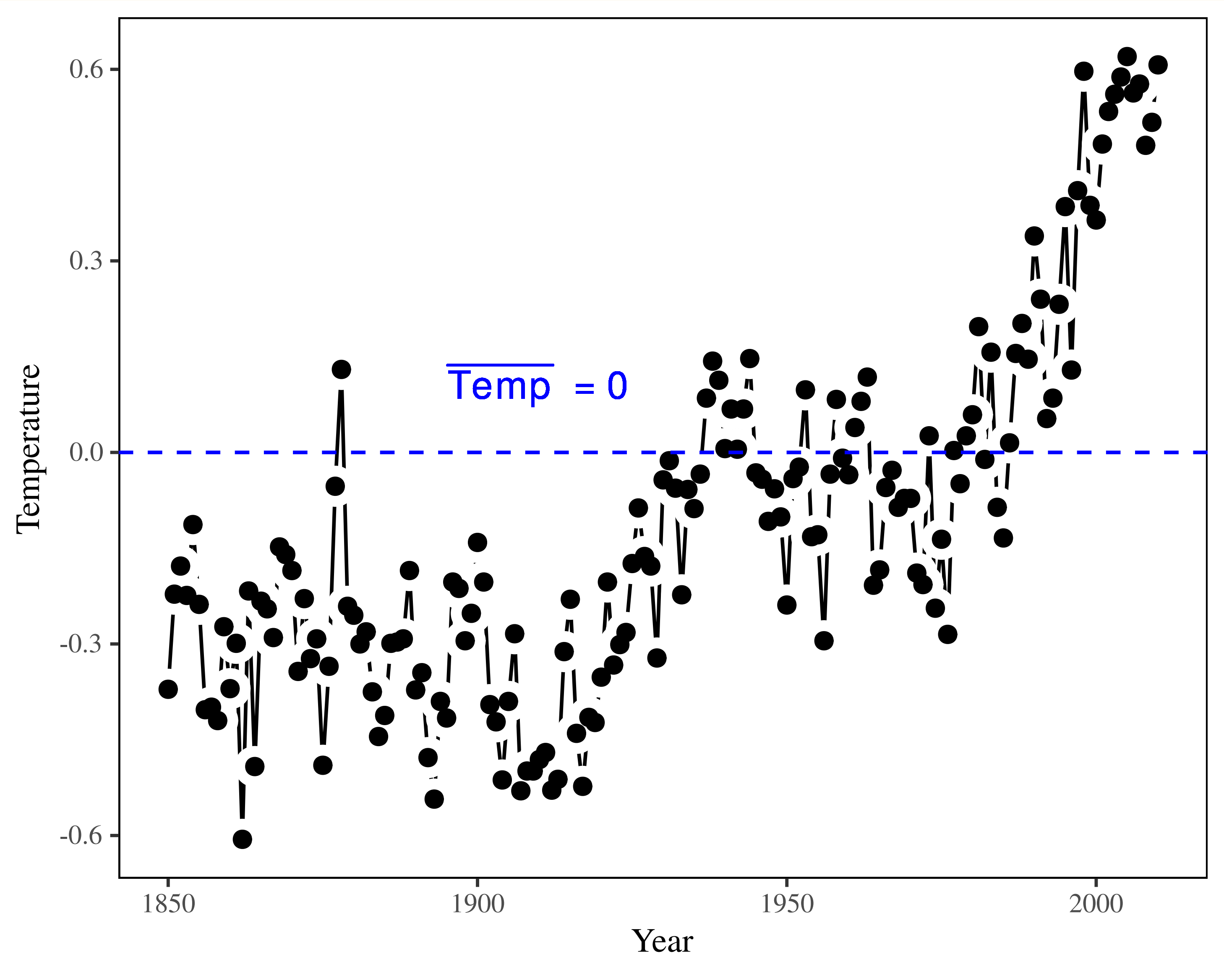
	Year	Temperature
1	1850	-0.371
2	1851	-0.222
3	1852	-0.178
4	1853	-0.224
5	1854	-0.113
6	1855	-0.238
7	1856	-0.403
8	1857	-0.399
9	1858	-0.420
10	1859	-0.273
11	1860	-0.370
12	1861	-0.299
13	1862	-0.606
14	1863	-0.217
15	1864	-0.492
16	1865	-0.233
17	1866	-0.245

- **Year:** year in which yearly average temperature was computed, from 1850 to 2010
- **Temperature:** northern hemisphere temperature minus the 161-year average (degrees Celsius)

Exploratory Data Analysis (EDA)



Exploratory Data Analysis (EDA)



Moving away from the independence assumption

- Up until now, we assumed responses are approximately normal and independent of one another
- The assumption of independence is rarely (if ever) quite right, but it is often a reasonable approximation
- We see presence of cluster effects and serial effects in many real life scenarios

Examples:

- a single unit of observation (person, organization, nation, etc.) is tracked over many time periods or points of time
- time periods that are close to one another are more likely to be similar than time periods that are relatively remote, e.g. stock prices, covid case counts

Time series with serial correlation

- Observed values will go on extended excursions away from the long-run mean
- Residuals exhibits **runs** i.e., consistently positive or negative for long periods
- Sample averages at different time segments do not estimate the correct or long-run mean

Standard Error of an Average in Time Series

$$SE(\bar{Y}) = \sqrt{\frac{1+r_1}{1-r_1}}\frac{s}{\sqrt{n}}$$

- sample variability will no longer be $s/\sqrt{n}$ (i.e., the sample standard error assuming the data is independent)
- r_1 is called the sample *first serial correlation coefficient*

First-order autoregressive model

The series, $\{Y_t\}$, is measured at equally spaced points in time

The deviation of an observation at time t from the long-run series mean v is $(Y_t - v)$

$\mu\{(Y_t - v) \mid \text{past history}\}$, the mean of the t^{th} deviation as a function of all previous deviations depends on lag 1 deviation

$$\mu\left\{(Y_t - v) \middle| \text{ past history }\right\} = \alpha\left(Y_{t-1} - v\right)$$

- α is called the autoregression coefficient

The first serial correlation coefficient, r_1

r_1 provides a numerical summary measure of the correlation between adjacent residuals

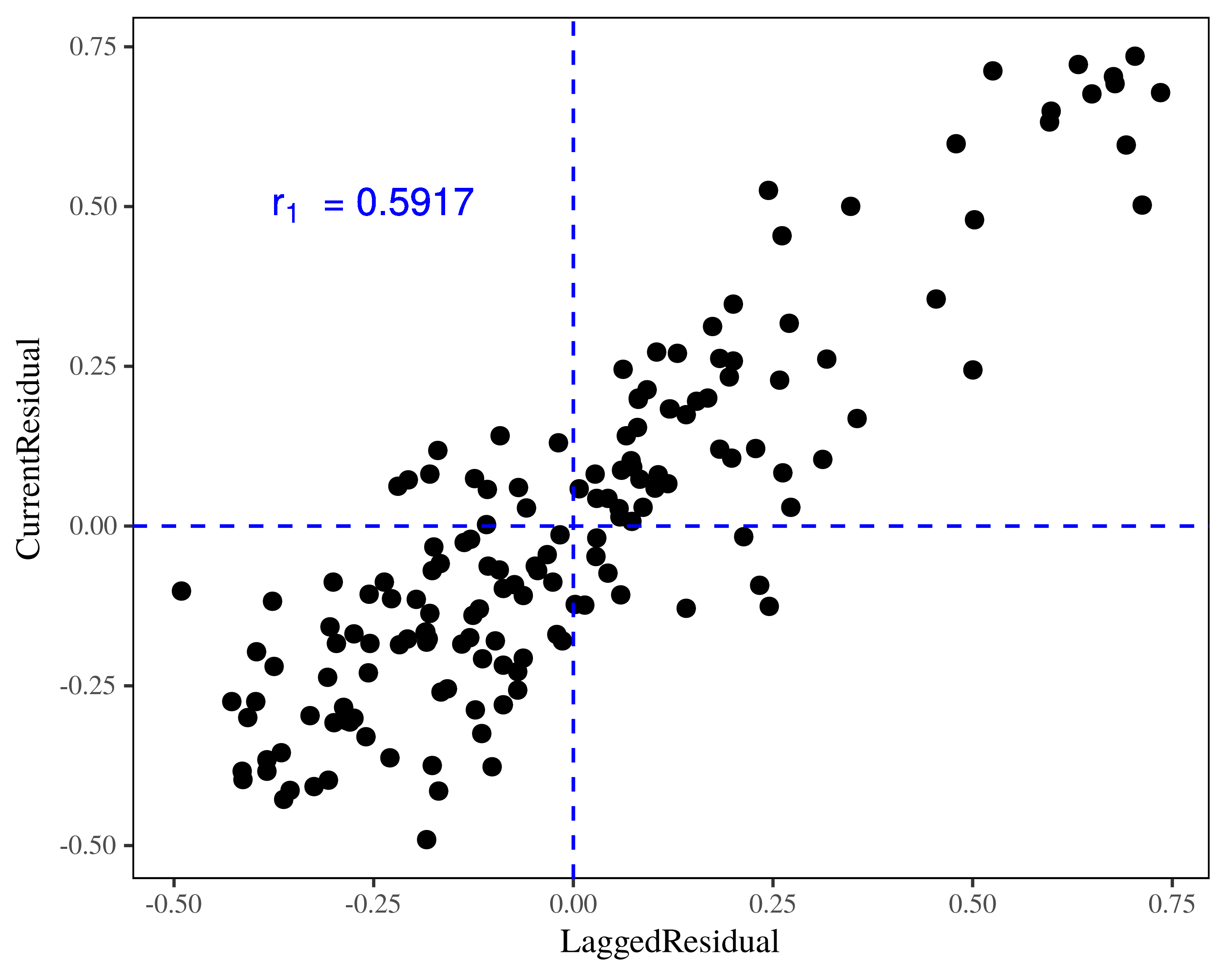
- r_1 is similar to the estimated slope in a regression of Y_t on Y_{t-1}

$$r_1 = \frac{c_1}{c_0}$$

$$c_1 = \tfrac{1}{n-1} \sum_{t=2}^{n} \text{res}_{\,t} \times \text{res}_{\,t-1} \qquad \text{and} \qquad c_0 = \tfrac{1}{n-1} \sum_{t=1}^{n} \text{res}_{\,t}^{\,2},$$

- c_0 is just the sample variance of the residuals
- c_1 is just the sample covariance of the residuals that are 1 lag apart

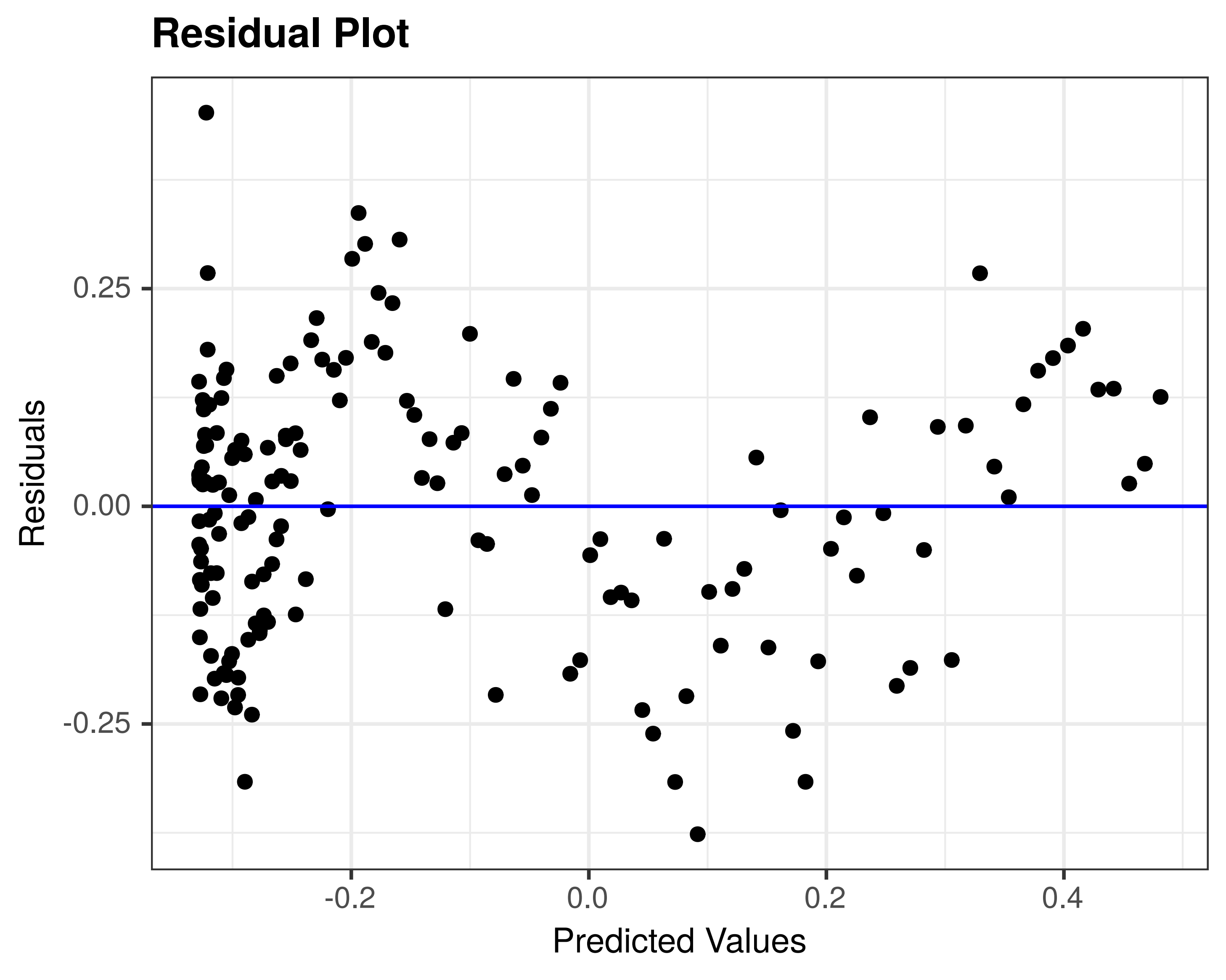
Temperature example: a lag plot



Temperature example: linear fit with a quadratic term

```
gwarming_m1 <- lm(Temperature ~ Year + yearSquared, data = case1502)
summary(gwarming_m1)

Call:
lm(formula = Temperature ~ Year + yearSquared, data = case1502)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.37663 -0.10532  0.01289  0.10494  0.45210 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.32095    0.01607 -19.966  < 2e-16 ***
Year         0.12594    0.04475   2.814  0.00551 ** 
yearSquared  0.54860    0.06131   8.948 9.29e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1503 on 158 degrees of freedom
Multiple R-squared:  0.7163,	Adjusted R-squared:  0.7127 
F-statistic: 199.5 on 2 and 158 DF,  p-value: < 2.2e-16
```

Temperature example: residual plot



Remedy: regression with filtered variables (AR(1) model)

Serial correlation can be handled by a special transformation called *filtering*

$$V_t = Y_t - \alpha Y_{t-1}, \quad \text{and} \quad U_t = X_t - \alpha X_{t-1}$$

The model reduces to the following with no more correlated errors and same slope

$$\mu\{V_t \mid U_t\} = \gamma_0 + \beta_1 U_t$$

- The intercept changes to $\gamma_0 = (1 - \alpha)\beta_0$.

Use same modeling concepts on the filtered response and explanatory variables to estimate the regression coefficients (Also applies to MLR!)

Remedy: regression with filtered variables (AR(1) model)

Problem: one must know α to construct the filtered variables V and U

Solution: Use r_1 can be used as an estimate of α

$$V_t = Y_t - r_1 Y_{t-1} \quad \text{and} \quad U_t = X_t - r_1 X_{t-1}$$

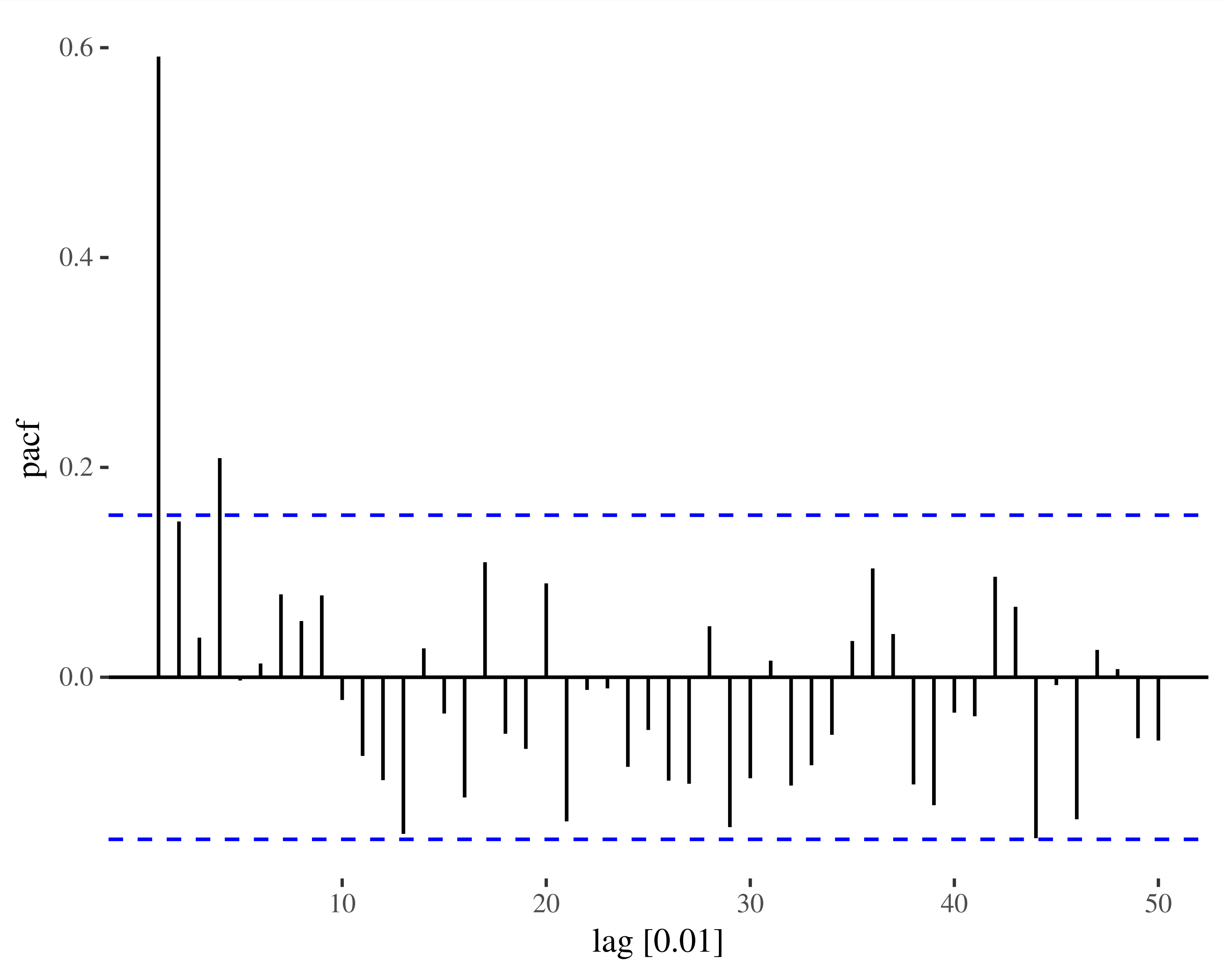
Partial autocorrelation

Partial autocorrelation measures the association between residuals spaced a certain number of lags apart in time, after accounting for the effects of the residuals between them.

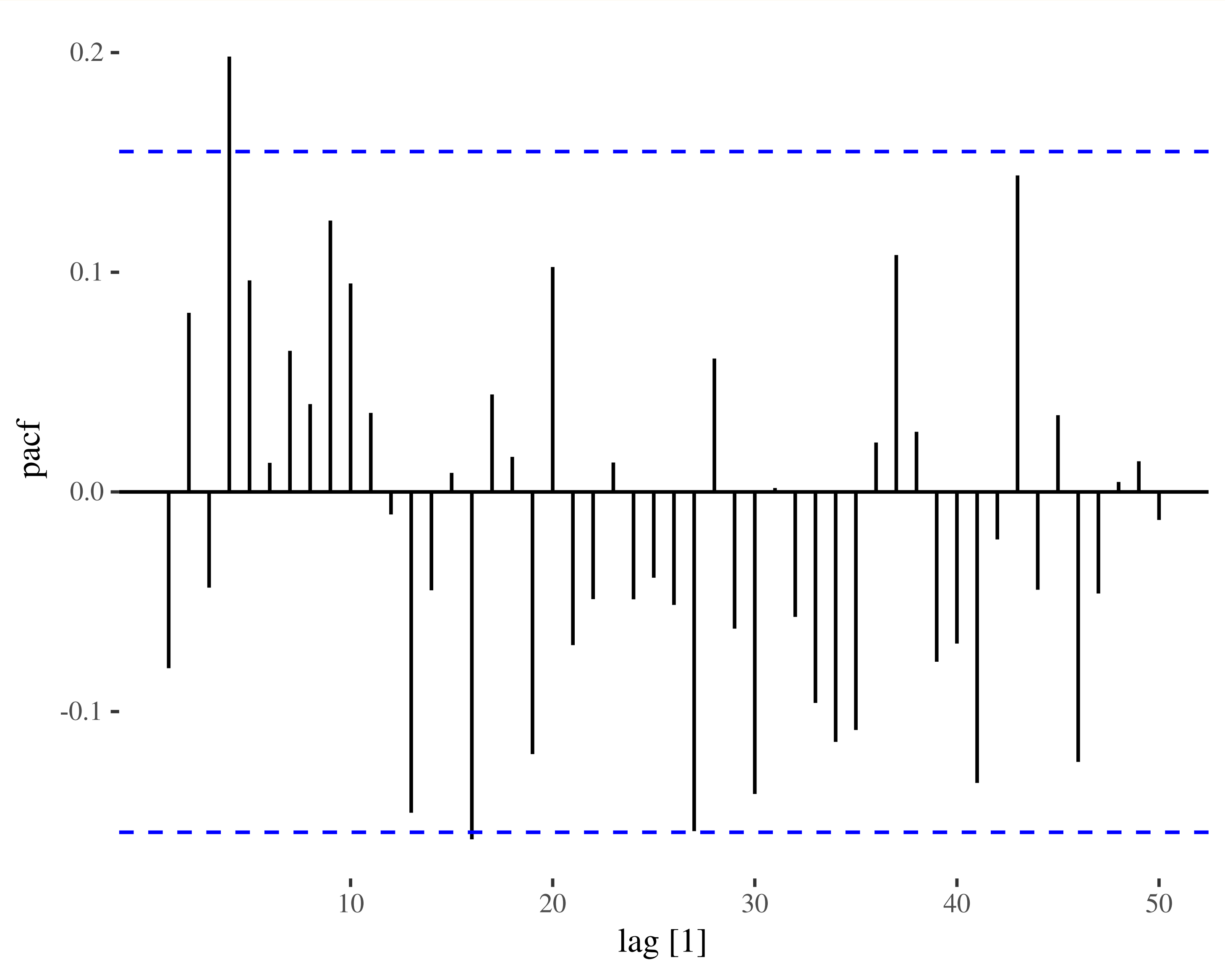
```
residual <- gwarming_m1$residuals
gwarming_Pacf <- pacf(residual, plot = FALSE)  # partial autocorrelation from residuals
```

```
r1 <- gwarming_Pacf$acf[1]  # First serial correlation coefficient
r1
[1] 0.5916607
```

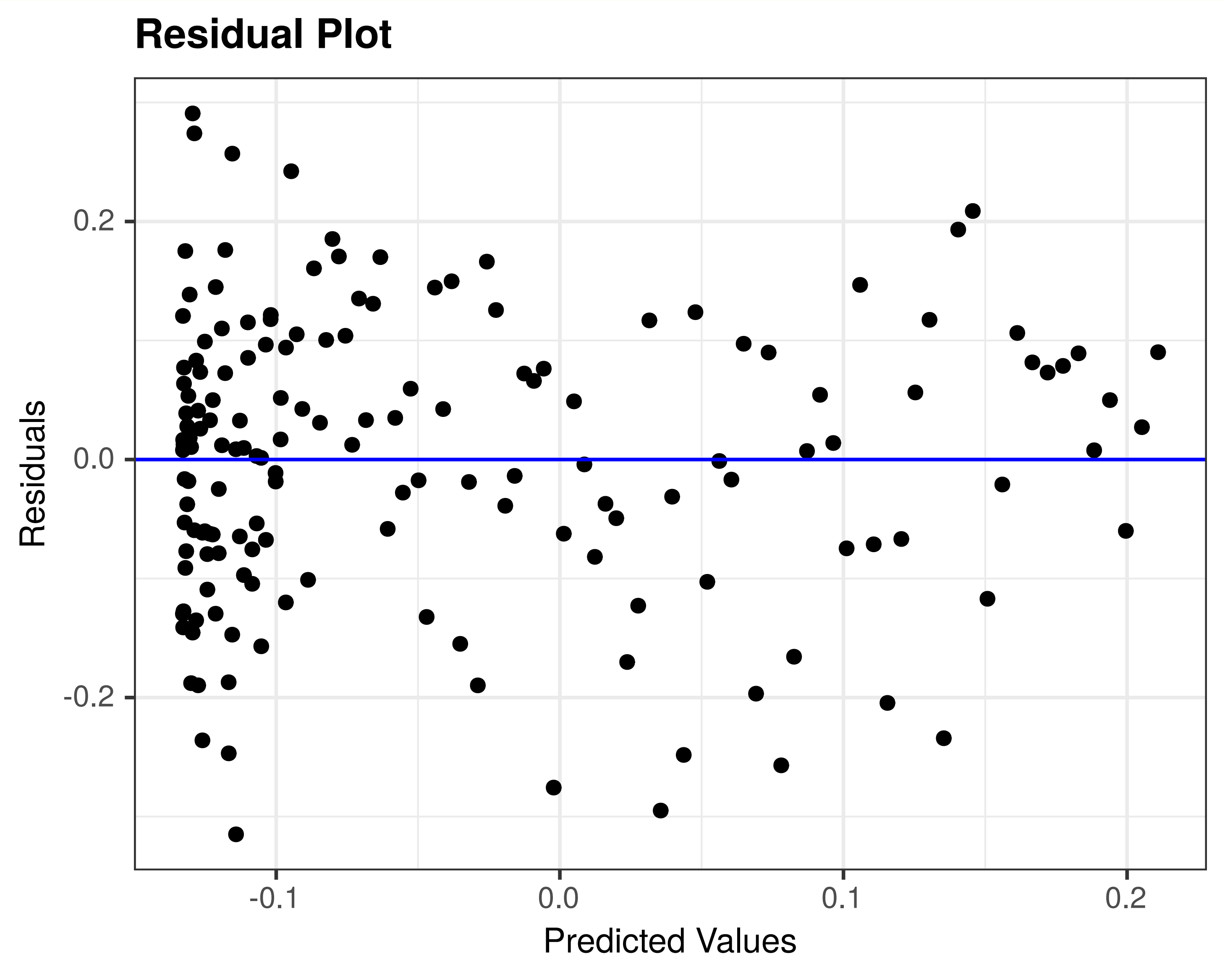
Partial autocorrelation functions



Refit the model on filtered variables



Residual plot after filtering

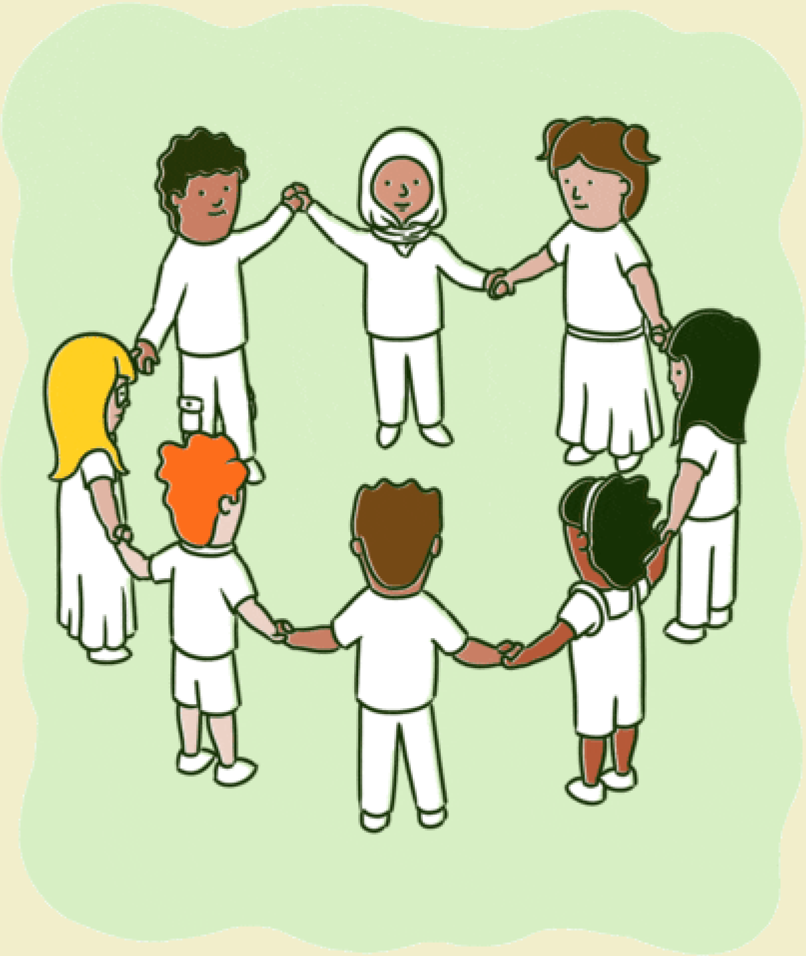


Procedure for filtering with more than one predictor

1. Fit the ordinary regression of the response on the explanatory variables and obtain residuals.
2. Calculate the autocovariance estimates c_0 and c_1 from the residuals. From these calculate the first serial correlation coefficient $r_1 = c_1/c_0$.
3. Compute the filtered versions of the response and explanatory variables.
4. Fit the regression of the filtered response on the filtered explanatory variables, and use the usual tools to make inferences about the coefficients (but not the intercept). The intercept for the model of interest, if desired, is estimated by the reported intercept estimate divided by $(1 - r_1)$.

Your Turn 1

05:00



- Go over to the in class activity file
- Complete the activity in your group