Fitting time series models

CIHR Course Week 2

Michael Law

michael.law@ubc.ca

GLOBAL HEALTH
RESEARCH CORE

Teaching Objectives

- Describe time series modeling

- Learn about key strategies
    - Plotting
    - Time trends
    - Seasonality
    - Autocorrelation

- Discuss two examples

Steps we will cover

1. Setup data
2. Visually inspect the data
3. Perform preliminary analysis
4. Check for and address autocorrelation
5. Run the final model and plot the results

Step 1: Setup Data

Data Setup

- One data row for each time period, including:
  - Time variable for day/month/year
  - Outcome of interest at that time period

- Possible variables:
  - Time Trend
  - Post-intervention Level Change
  - Post-intervention Trend Change
  - Outcome of interest

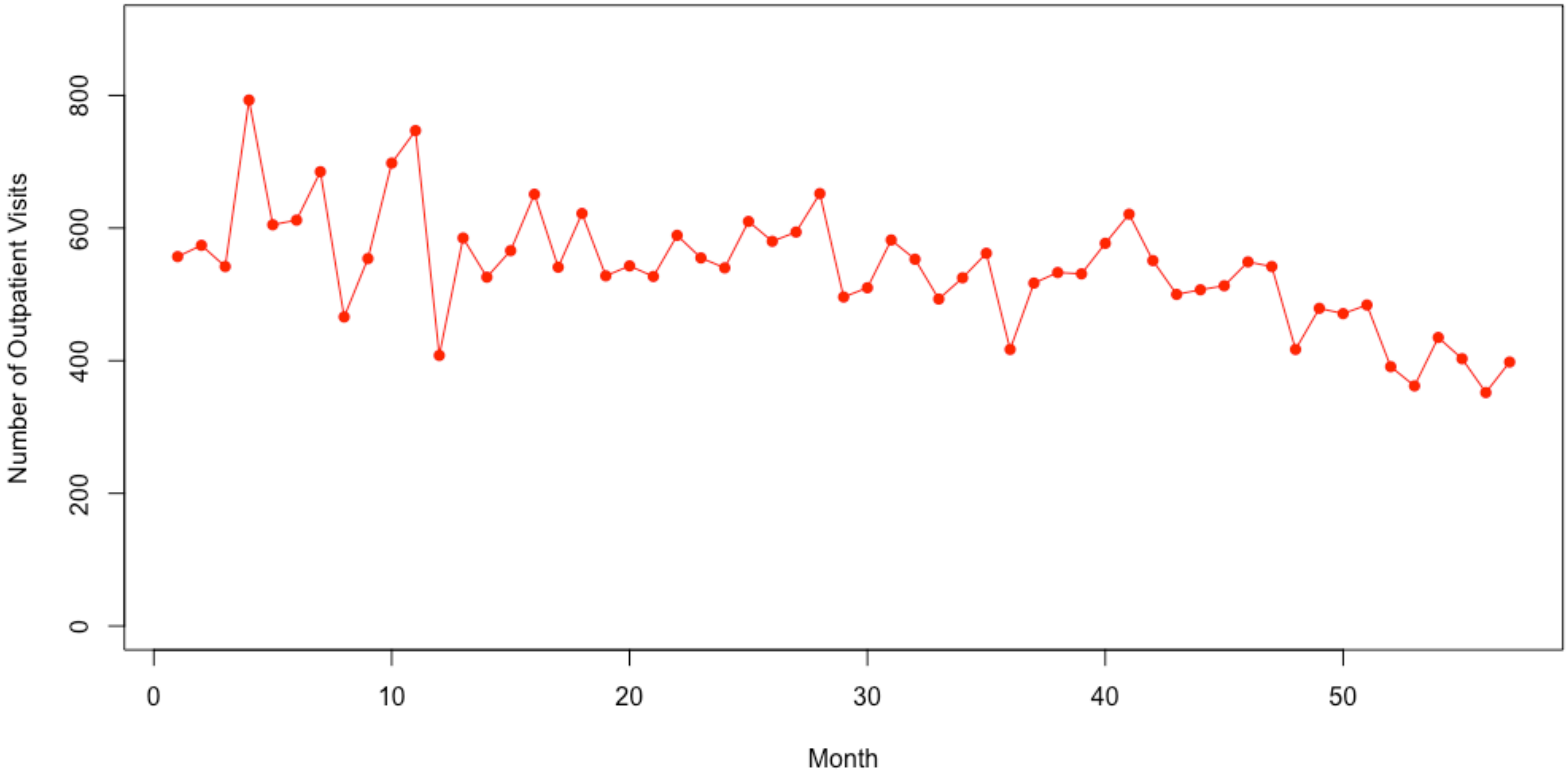
Month	Visits
2016-01-01	557
2016-02-01	574
2016-03-01	542
2016-04-01	793
2016-05-01	605
2016-06-01	612
2016-07-01	685
2016-08-01	466
2016-09-01	554
2016-10-01	698
…	…

Step 2: Visually Inspect Data

Visually Inspect Data

What you are looking for:

- Trends
  - Linear
  - Polynomial
  - Seasonal (next week!)
- Potential interventions
- Data quality issues
  - Missing data
  - "Wild" points


Number of Outpatient Visits
0
200
400
600
800
0
10
20
30
40
50
Month

Step 3: Preliminary Analysis

Where we are now

1. Setup data
2. Visually inspect the data
3. Preliminary analysis
4. Check for and address autocorrelation
5. Run the final model and plot the results

Preliminary analysis

- Run a standard OLS regression with a time series specification

- We will work through:
  - Intercept-only
  - Simple time trend
  - Quadratic time trend

- This will form the basis for checks about autocorrelation

Intercept-only Model

$$outcome_t = \beta_0 + \varepsilon_t$$

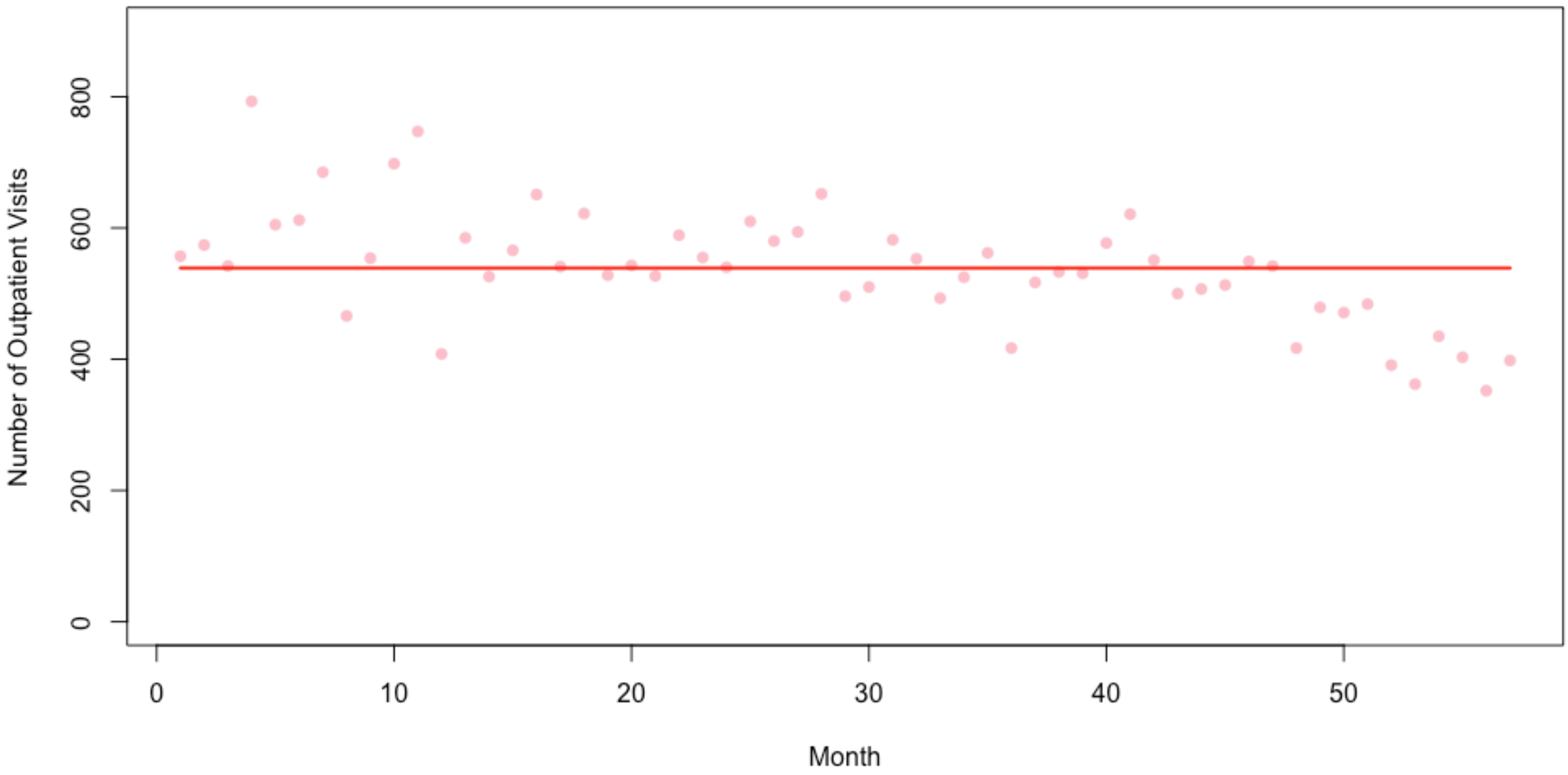
```
> summary(ols.model)

Call:
lm(formula = count ~ 1, data = dataset)

Residuals:
     Min       1Q   Median       3Q      Max 
-186.965  -42.965    3.035   43.035  254.035 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   538.96      11.67   46.19   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 88.09 on 56 degrees of freedom
```



Number of Outpatient Visits
0
200
400
600
800
0
10
20
30
40
50
Month

Intercept and Time Model

$$outcome_t = \beta_0 + \varepsilon_t$$

↓

$$outcome_t = \beta_0 + \beta_1 \cdot time + \varepsilon_t$$

GLOBAL HEALTH RESEARCH CORE

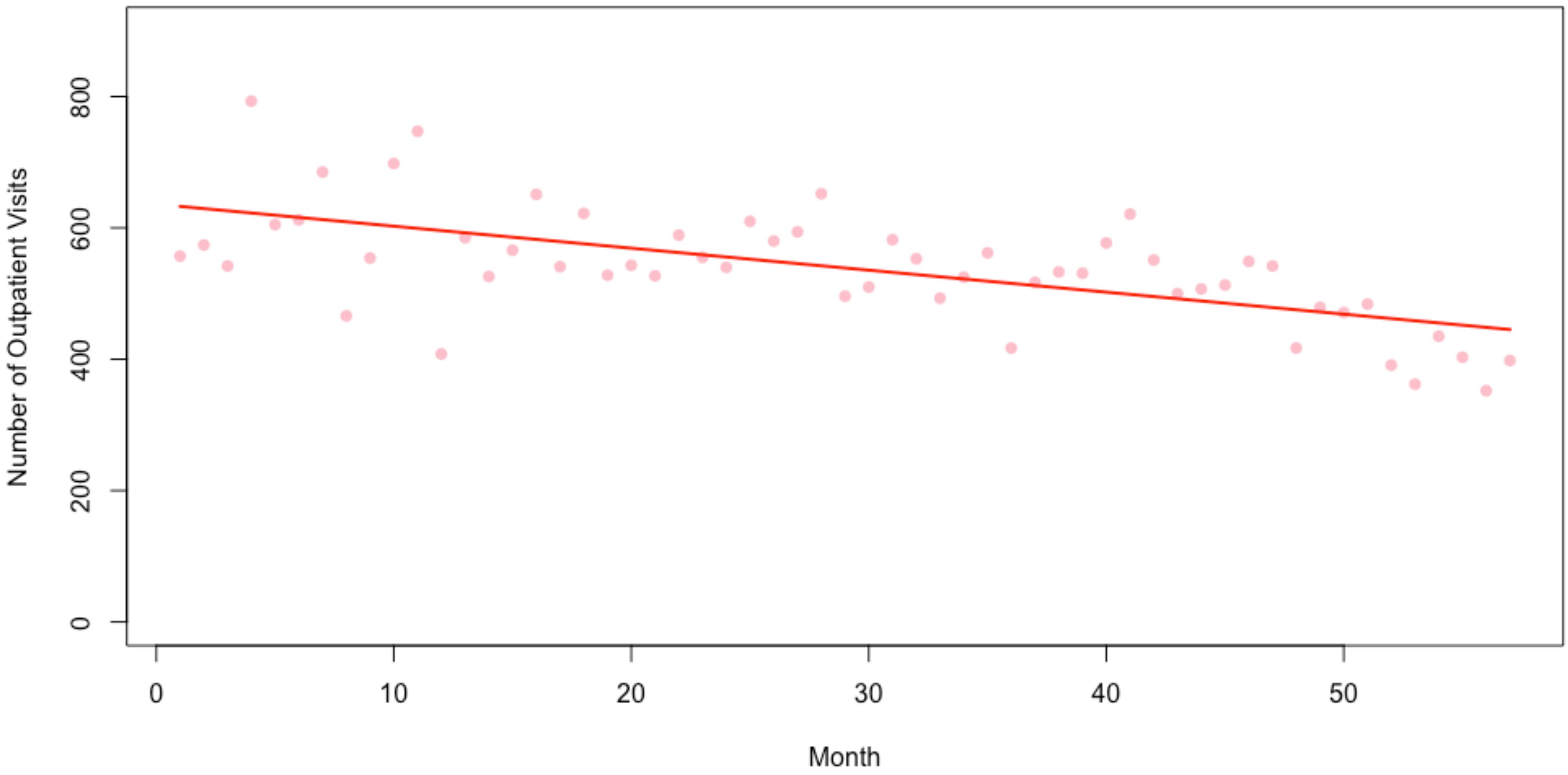
```
> summary(ols.model.time)

Call:
lm(formula = count ~ time, data = dataset)

Residuals:
     Min       1Q   Median       3Q      Max 
-187.891  -44.451    2.355   46.201  170.320 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 636.0739    18.5113  34.361  < 2e-16 ***
time         -3.3486     0.5552  -6.031 1.43e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 68.96 on 55 degrees of freedom
Multiple R-squared:  0.3981,	Adjusted R-squared:  0.3872 
F-statistic: 36.38 on 1 and 55 DF,  p-value: 1.432e-07
```



GLOBAL HEALTH
RESEARCH CORE

Intercept and Time Model

$$outcome_t = \beta_0 + \varepsilon_t$$

$$outcome_t = \beta_0 + \beta_1 \cdot time + \varepsilon_t$$

$$outcome_t = \beta_0 + \beta_1 \cdot time + \beta_2 \cdot time^2 + \varepsilon_t$$

GLOBAL HEALTH
RESEARCH CORE

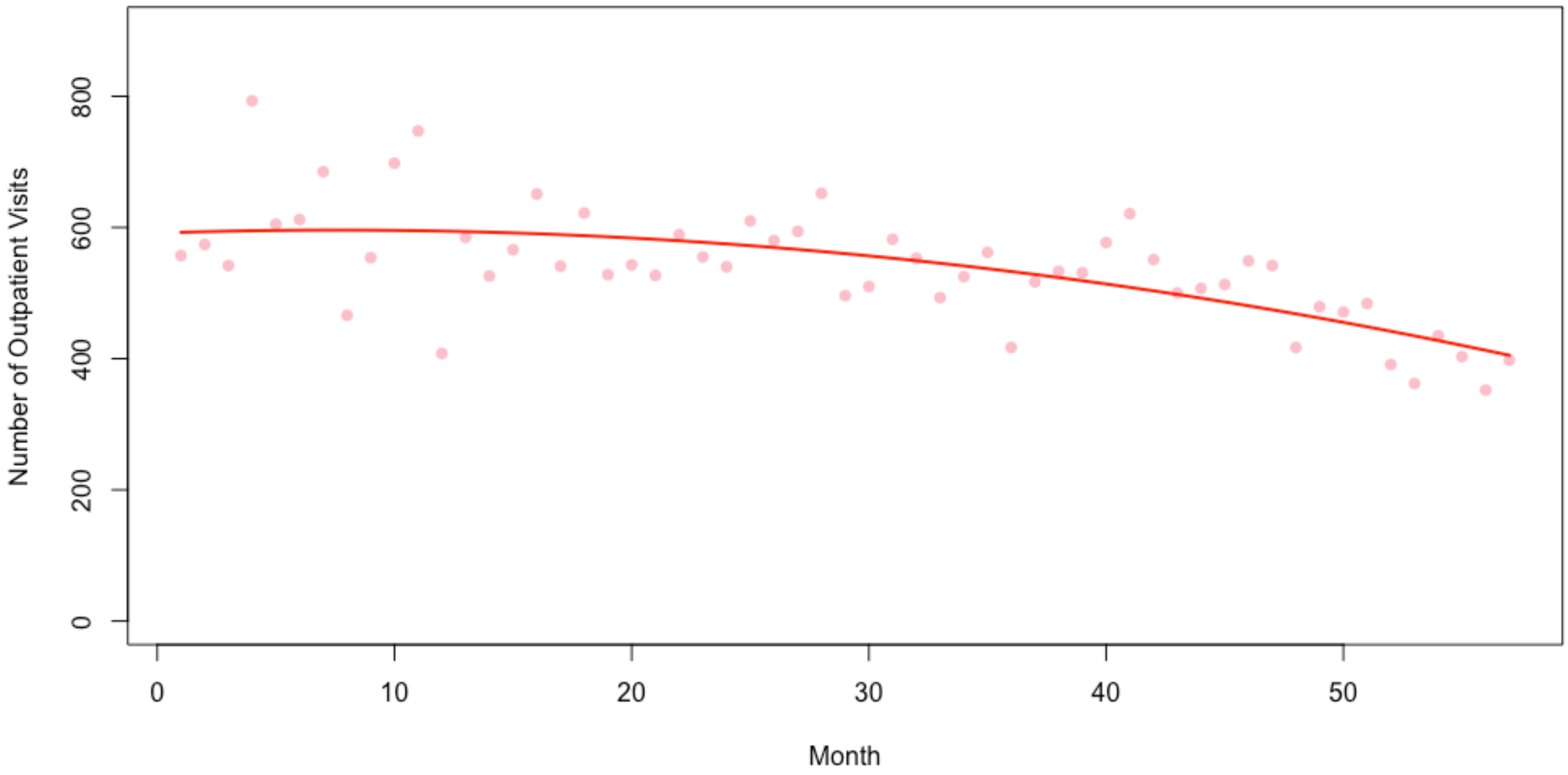
```
> summary(ols.model.time2)

Call:
lm(formula = count ~ time + time2, data = dataset)

Residuals:
     Min       1Q   Median       3Q      Max 
-186.458  -46.691    3.632   28.892  198.012 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 591.50239   27.51523  21.497   <2e-16 ***
time          1.18411    2.18884   0.541   0.5907    
time2        -0.07815    0.03658  -2.136   0.0372 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 66.83 on 54 degrees of freedom
Multiple R-squared:  0.445,	Adjusted R-squared:  0.4244 
F-statistic: 21.65 on 2 and 54 DF,  p-value: 1.247e-07
```



GLOBAL HEALTH RESEARCH CORE

Step 4: Autocorrelation

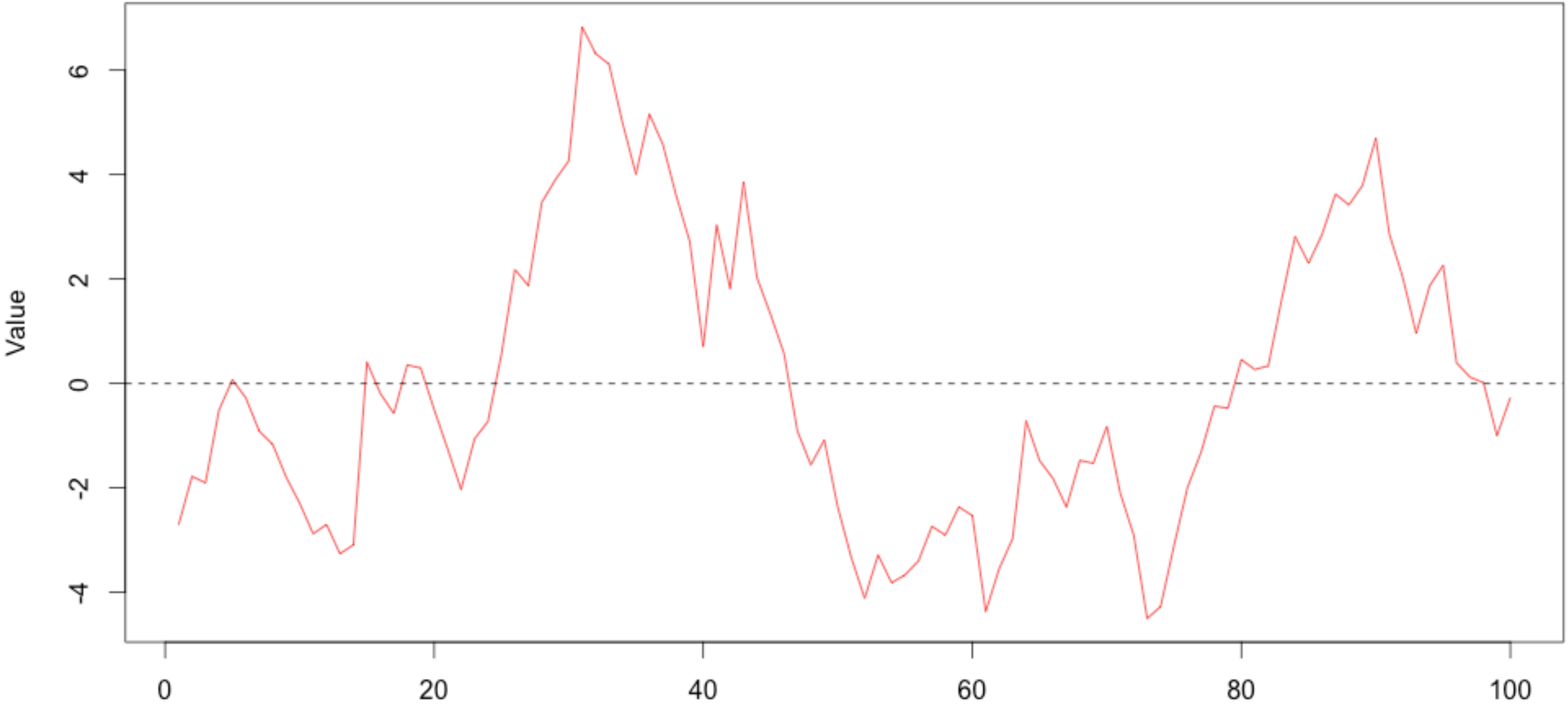
What is autocorrelation?

- Relationship in data points over time
  - Means the data points are not independent

- Two types we will discuss:
  - Autoregression
  - Moving Average

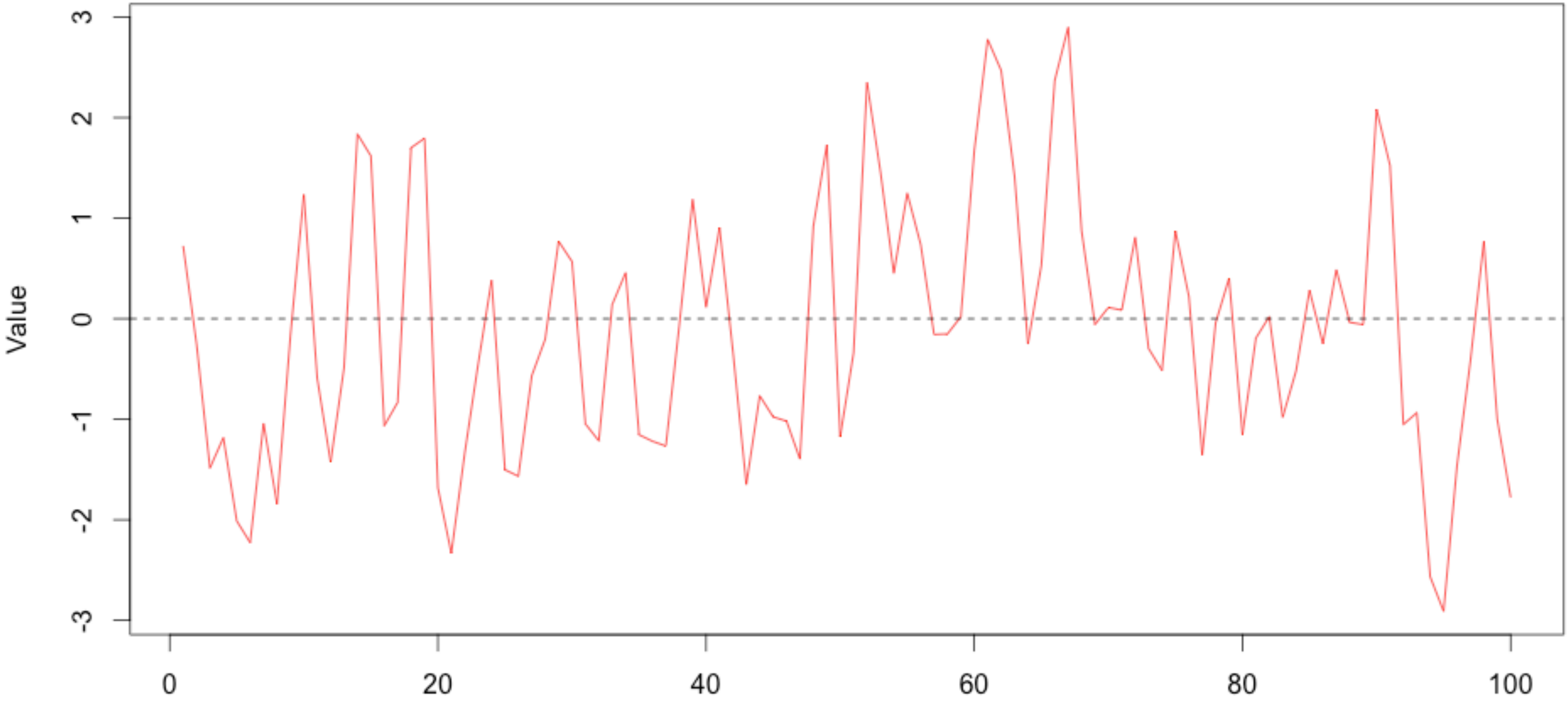
Autoregression

$$\varepsilon_t = \phi\varepsilon_{t-1} + \upsilon_t$$



Moving Average

$$\varepsilon_t = \upsilon_t + \vartheta \upsilon_{t-1}$$



Methods to Check Autocorrelation

- Several methods, including:
  - Durbin-Watson test
  - Residual plots
  - ACF and partial-ACF plots

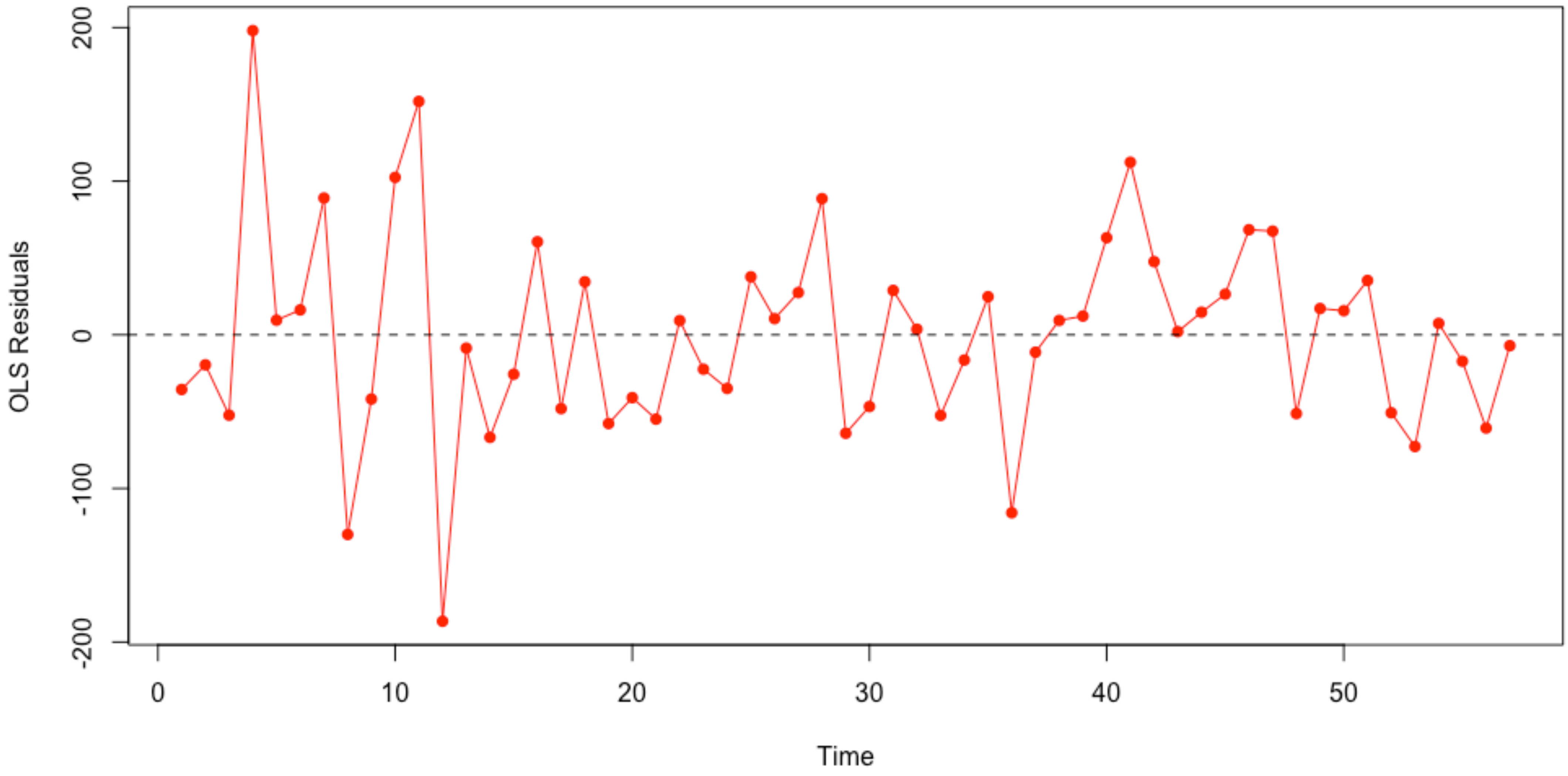
Durbin-Watson test

- A formal test that tests for correlated residuals
- Interpretation
  - Values of 2 indicate no autocorrelation
  - lower values indicate positive correlation, higher indicates negative correlation

```
> # Durbin-watson test, 8 time periods
> dwt(ols.model.time2,max.lag=12,alternative="two.sided")
 lag Autocorrelation D-W Statistic p-value
   1     -0.04948719      2.093509   0.960
   2     -0.13944970      2.256553   0.362
   3      0.07000156      1.825057   0.522
   4     -0.02657831      1.855414   0.730
   5     -0.02114350      1.822230   0.718
   6      0.10264861      1.562874   0.252
   7      0.13489035      1.460330   0.158
   8     -0.32012733      2.299303   0.060
   9      0.02045164      1.609675   0.596
  10     -0.14917289      1.894482   0.454
  11     -0.06803985      1.617629   0.826
  12      0.08589470      1.146209   0.062
 Alternative hypothesis: rho[lag] != 0
```

Residual plot

- Use a residual plot to visually inspect for independence



OLS Residuals
200
100
0
-100
-200
0
10
20
30
40
50
Time

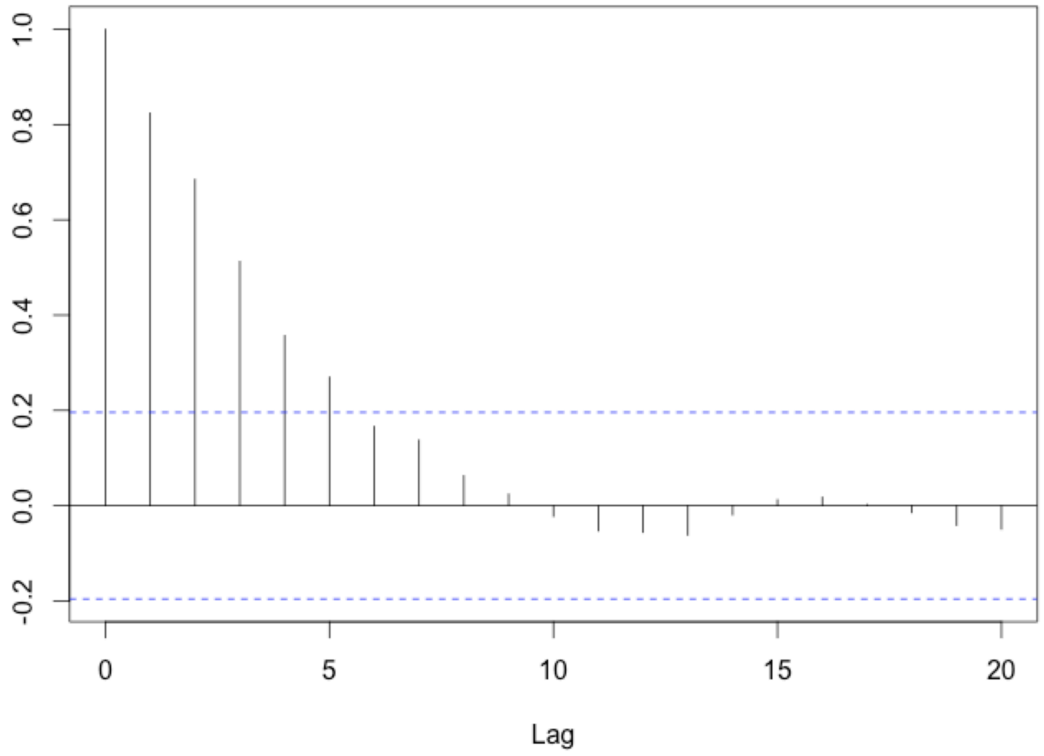
Autocorrelation plots

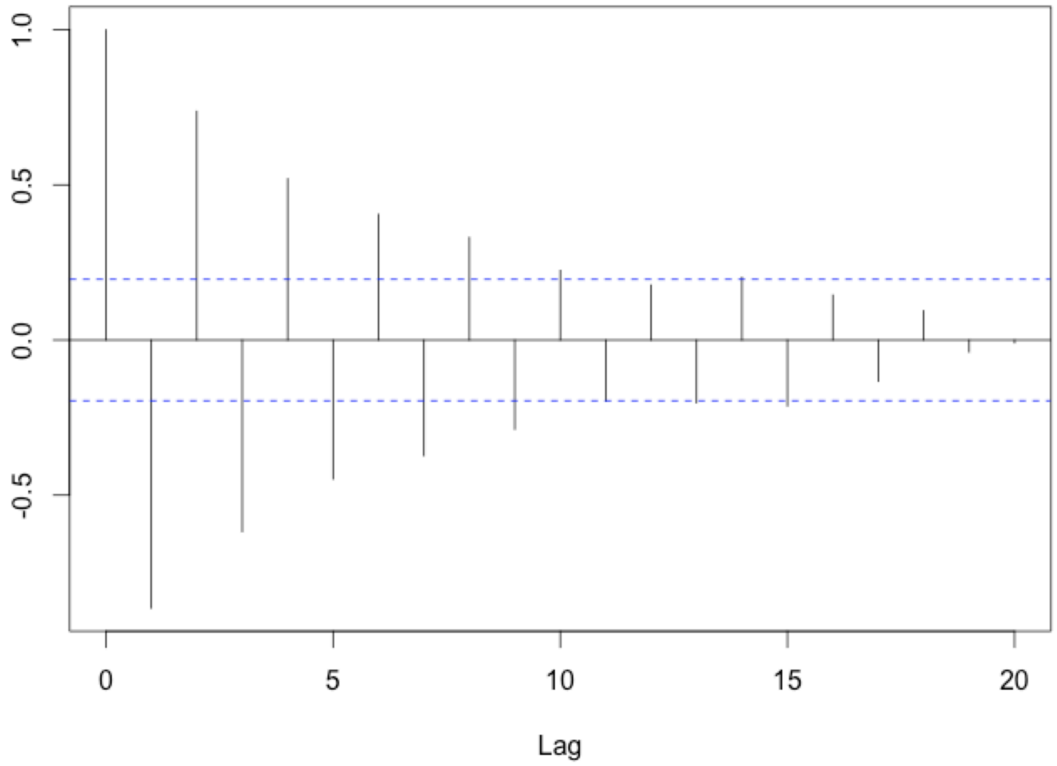
- A plotting method with which you can assess autocorrelation and moving averages

- Two plots
  - Autocorrelation
  - Partial autocorrelation

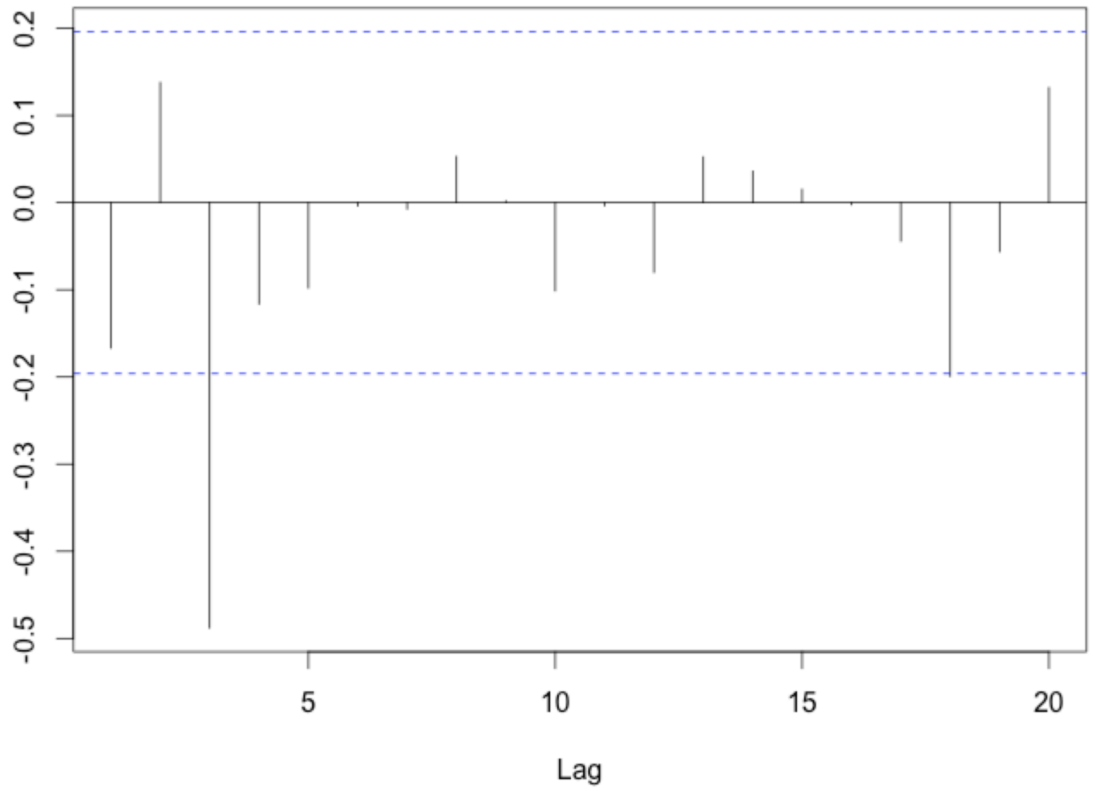
Model	ACF	Partial ACF
No autocorrelation	All zeros	All zeros
Autoregressive (p)	Exponential Decay	p significant lags before dropping to zero
Moving Average (q)	q significant lags before dropping to zero	Exponential Decay
Both (p,q)	Decay after q^{th} lag	Decay after p^{th} lag

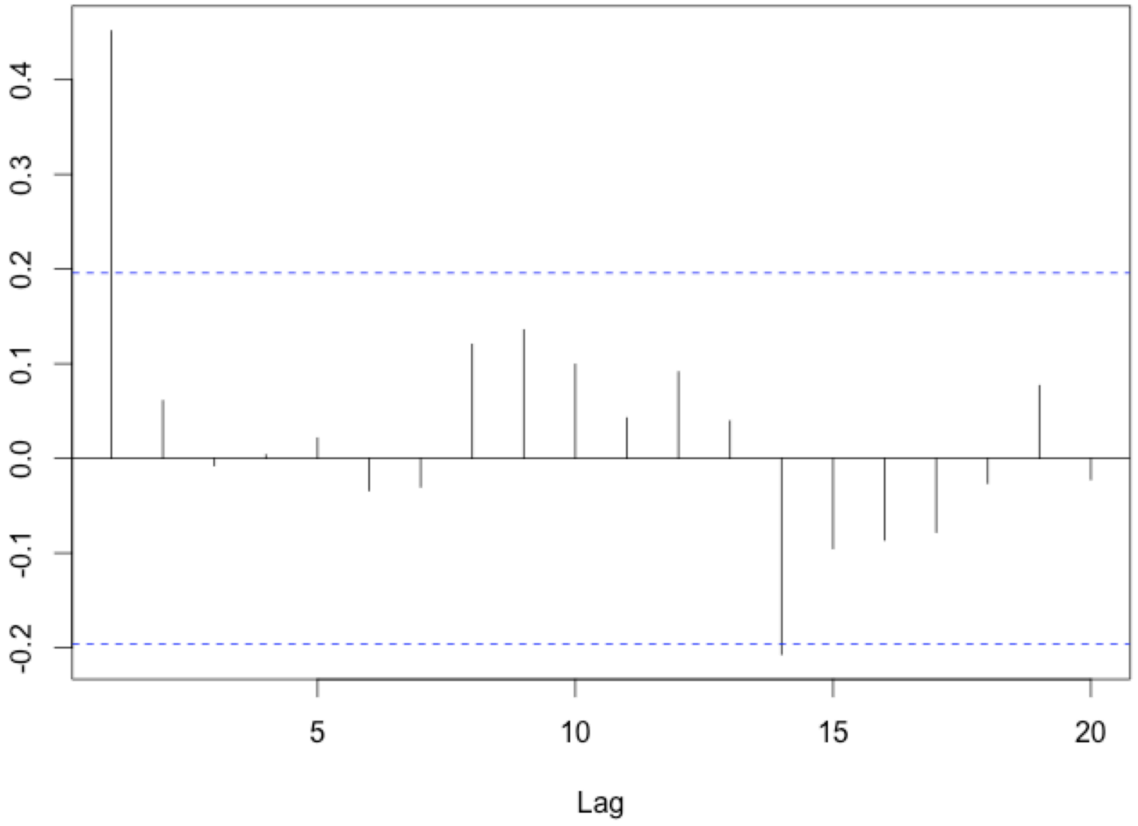
Examples of Exponential Decay

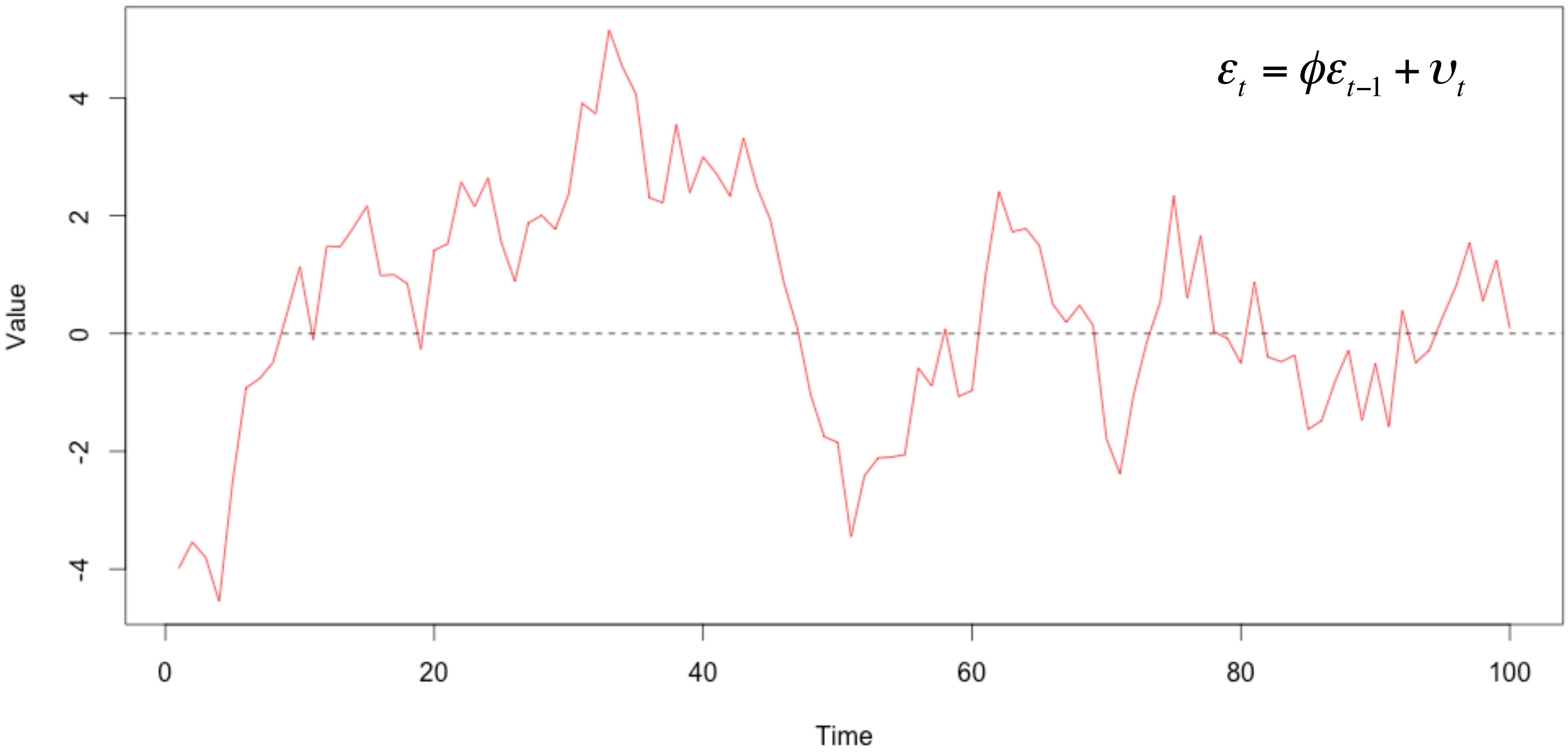


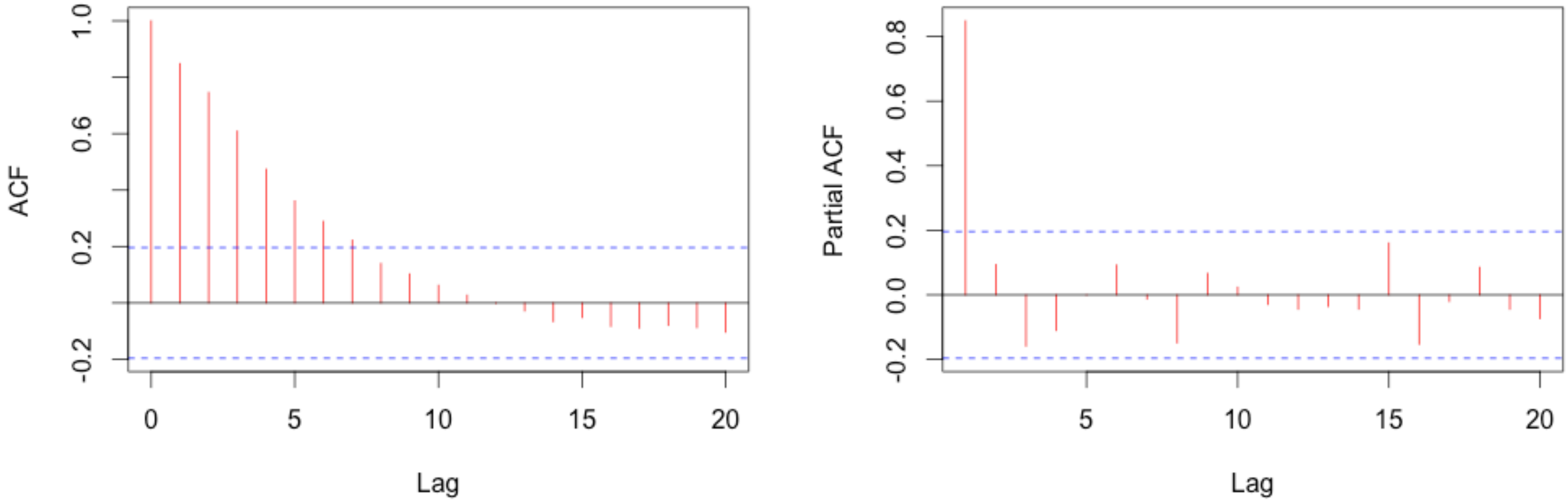


Examples of Significant Lags



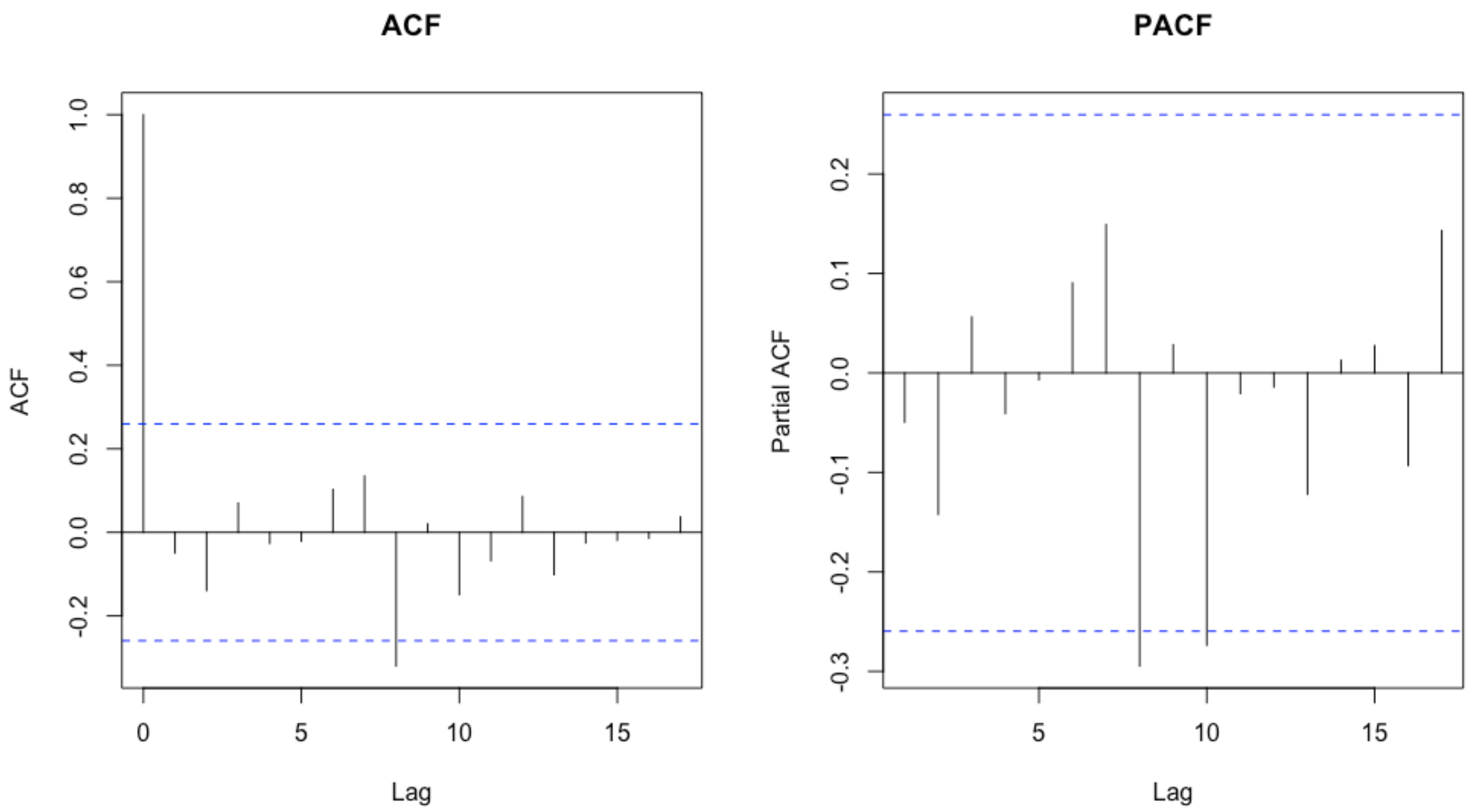



$\varepsilon_t = \phi\varepsilon_{t-1} + \upsilon_t$
Value
Time
4
2
0
-2
-4
0
20
40
60
80
100



ACF
1.0
0.6
0.2
-0.2
0
5
10
15
20
Lag
Partial ACF
0.8
0.6
0.4
0.2
0.0
-0.2
5
10
15
20
Lag

Model	ACF	Partial ACF
No autocorrelation	All zeros	All zeros
Autoregressive (p)	Exponential Decay	p significant lags before dropping to zero
Moving Average (q)	q significant lags before dropping to zero	Exponential Decay
Both (p,q)	Decay after q^{th} lag	Decay after p^{th} lag



GLOBAL HEALTH RESEARCH CORE

Step 5: Run the Final Model

Where we are now

1. Setup data
2. Visually inspect the data
3. Preliminary analysis
4. Check for and address autocorrelation
5. Run the final model and plot the results

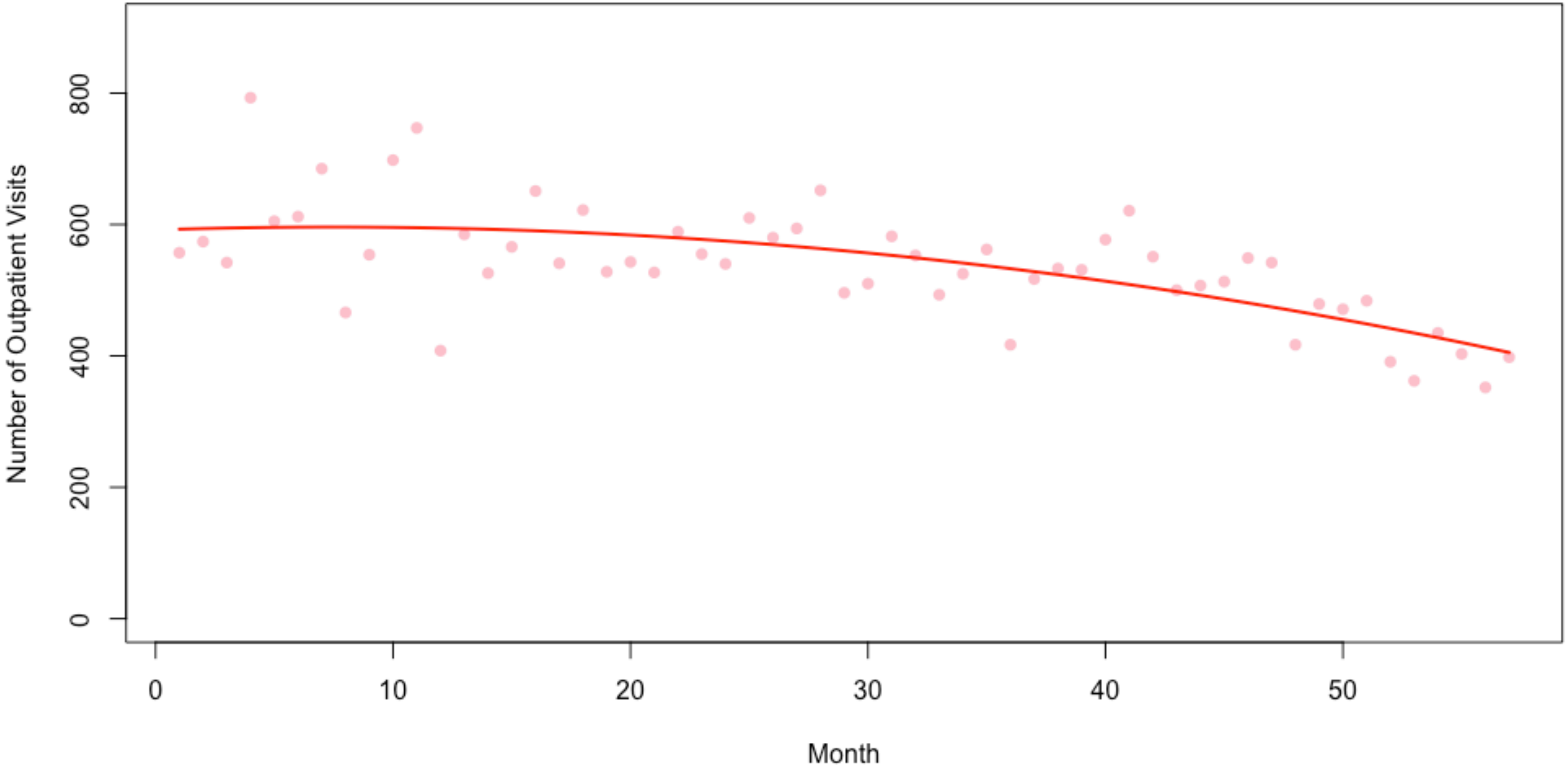
Running the final model

- Use function *gls* (similar specification to *lm*)
- If we wanted to use AR(1), for example

```
> summary(model_final)
Generalized least squares fit by maximum likelihood
  Model: count ~ time + time2
  Data: dataset
       AIC      BIC    logLik
  647.5837 657.799 -318.7919

Correlation Structure: AR(1)
 Formula: ~time
 Parameter estimate(s):
        Phi
-0.04894649

Coefficients:
               Value Std.Error   t-value p-value
(Intercept) 591.8056 26.306115 22.496884  0.0000
time          1.1618  2.092482  0.555205  0.5810
time2        -0.0778  0.034974 -2.224745  0.0303
```


Number of Outpatient Visits
0
200
400
600
800
0
10
20
30
40
50
Month

Next week…

1. Seasonality
2. Predictions
3. Syndromic surveillance models