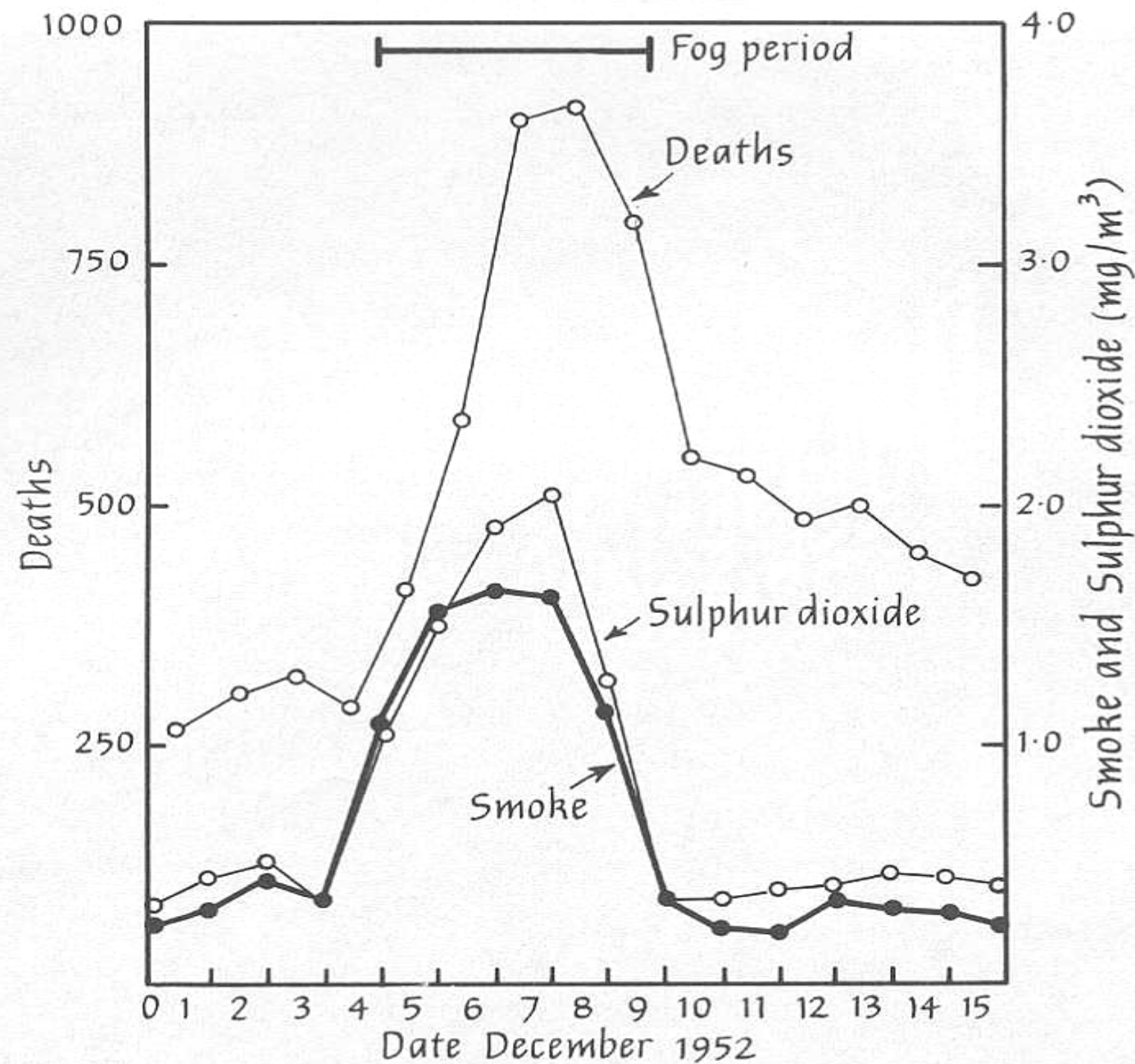


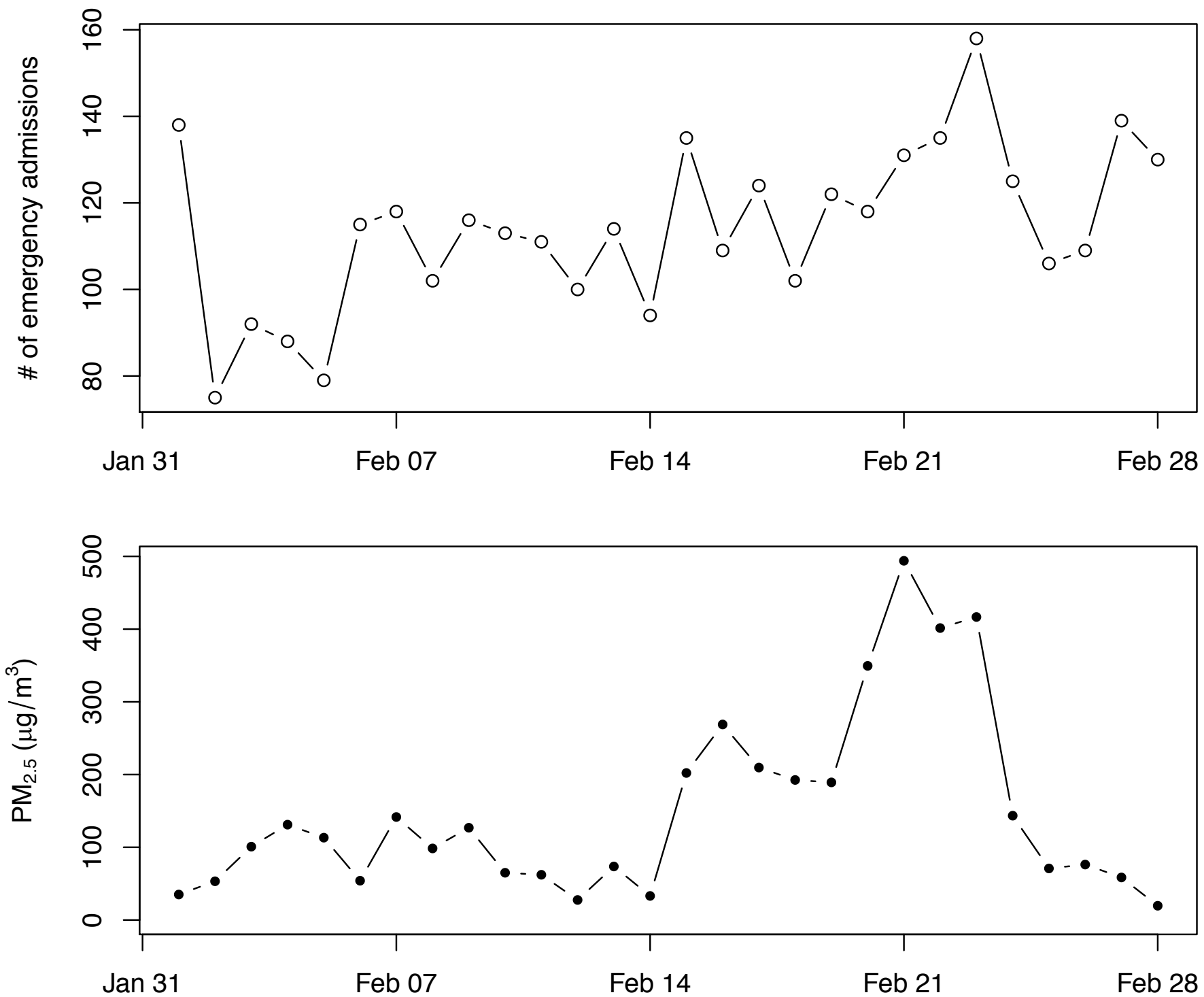
# Time Series Analysis

Biostatistics 140.712

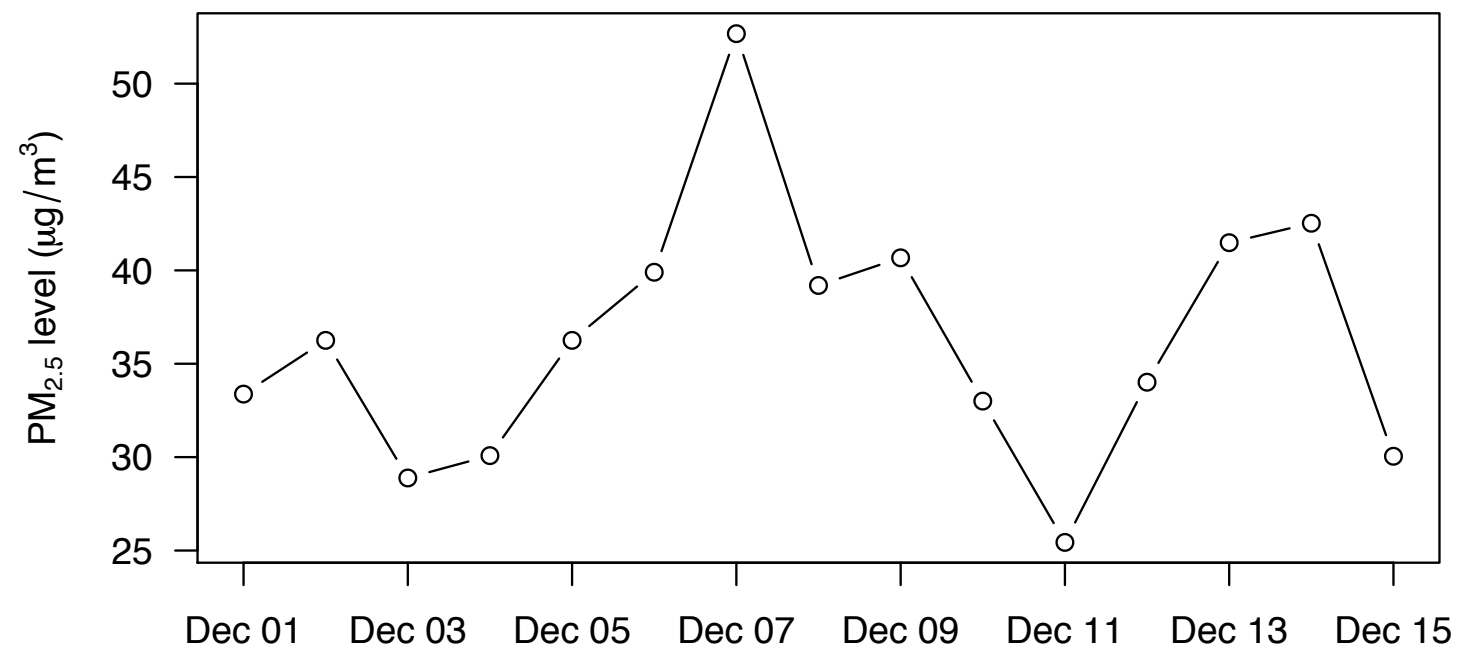
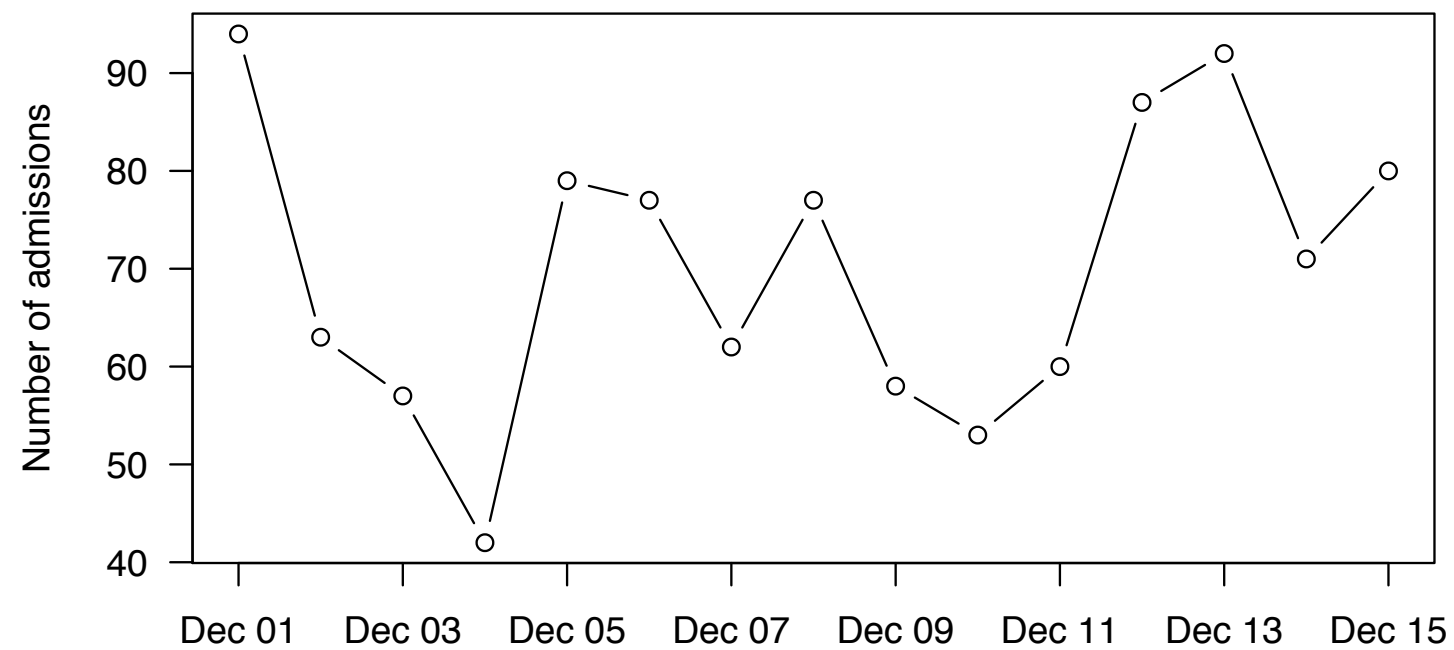
# London Fog (1952)



# Beijing Fog (2011)



# Chicago (2005)



# Time Series Analysis

$$Y_t \sim X_t \mid Z_t$$

- Relating changes in an exposure  $X$  with changes in an outcome  $Y$ , adjusting for potential confounders  $Z$
- The *units* of analysis are time points (seconds, minutes, days, months, etc.)
- Time series analysis is interesting because it comes with a “built-in covariate”: **Time itself**

# Time Series or Longitudinal?

- Time series
  - focuses on a single series of data
  - only has “within-subject” variation
- Longitudinal data
  - usually focuses on replicates of very short series across subjects
  - within- and between-subject variation
  - Very long time series replicated across many subjects sometimes called *functional data*
- No clear rule!

# Time is Special

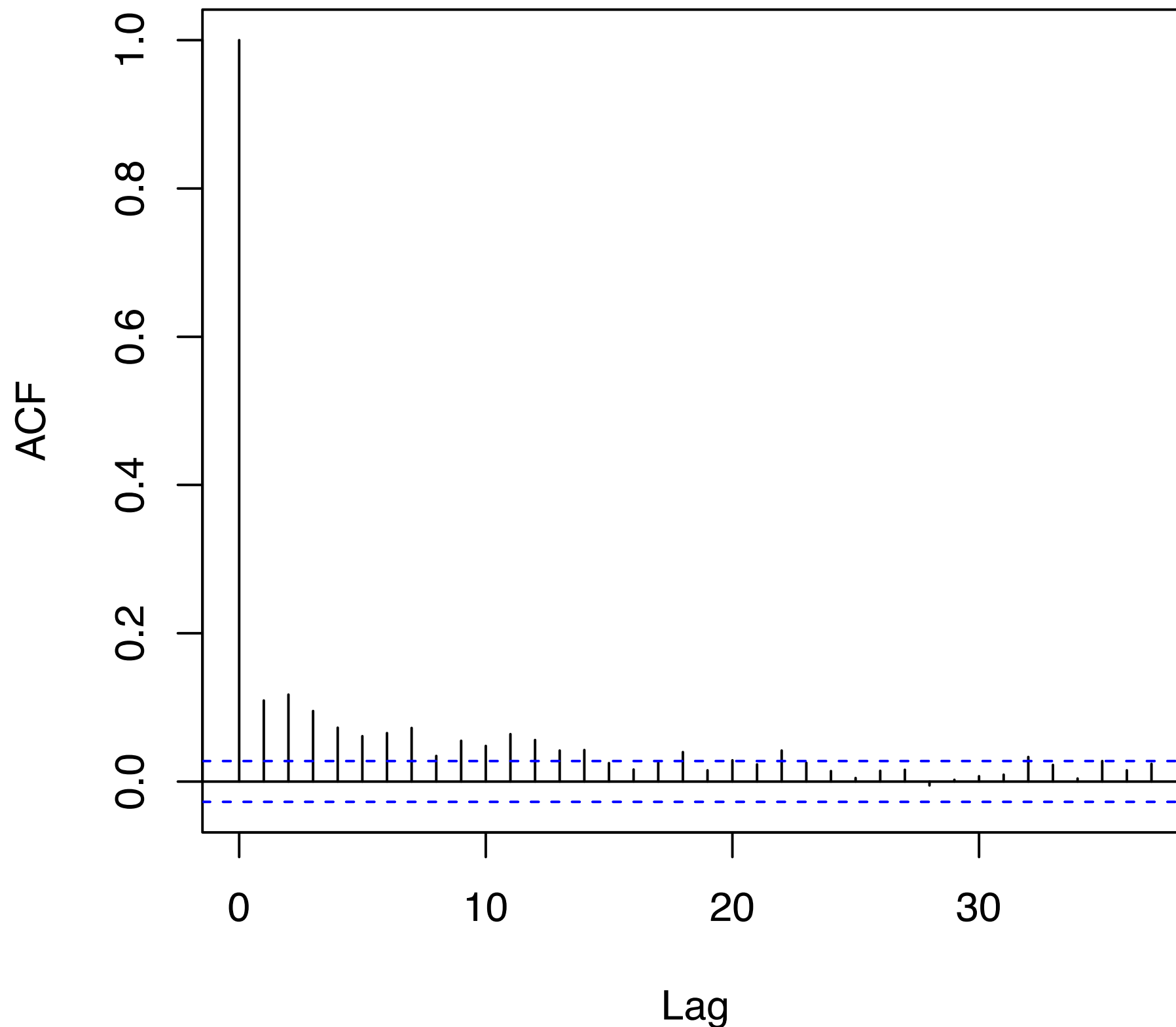
- Observations at neighboring time points are thought to be (positively) correlated with each other: **autocorrelation**
- Why are observations observed across time correlated?
- Time can be used as a “stand in” for other, **unobserved**, time-varying predictors
- Data can also be generated through a **dynamical process** whereby values at one time point may causally effect values at a future time point

# Autocorrelation

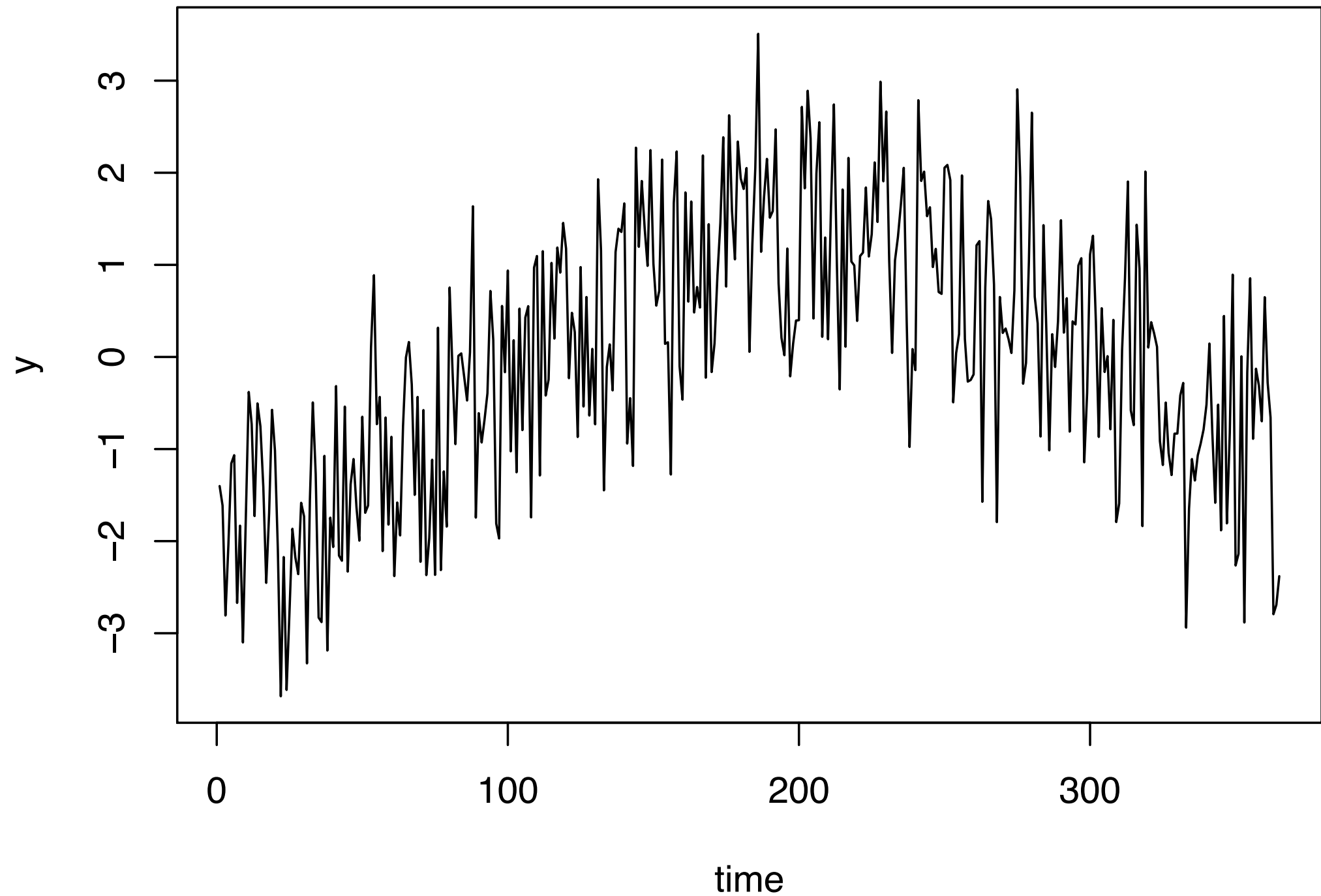
- Correlation between elements of a time series is called autocorrelation
- Autocorrelation can occur between elements of a time series at different lag distances
- Autocorrelation can be estimated using the autocorrelation function (acf)
- The acf can be plotted as a function of the lag distance



# Autocorrelation



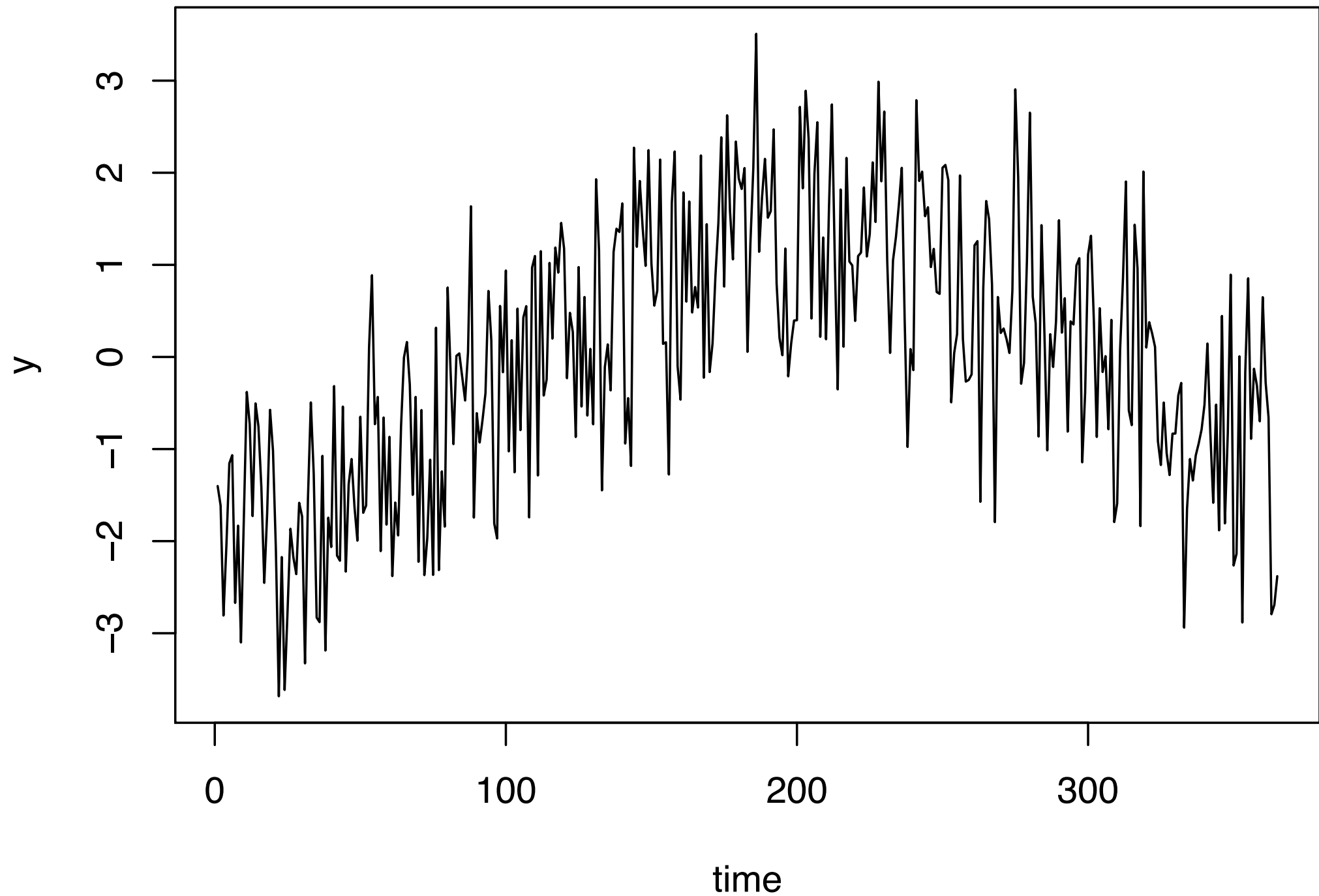
# Autocorrelation?



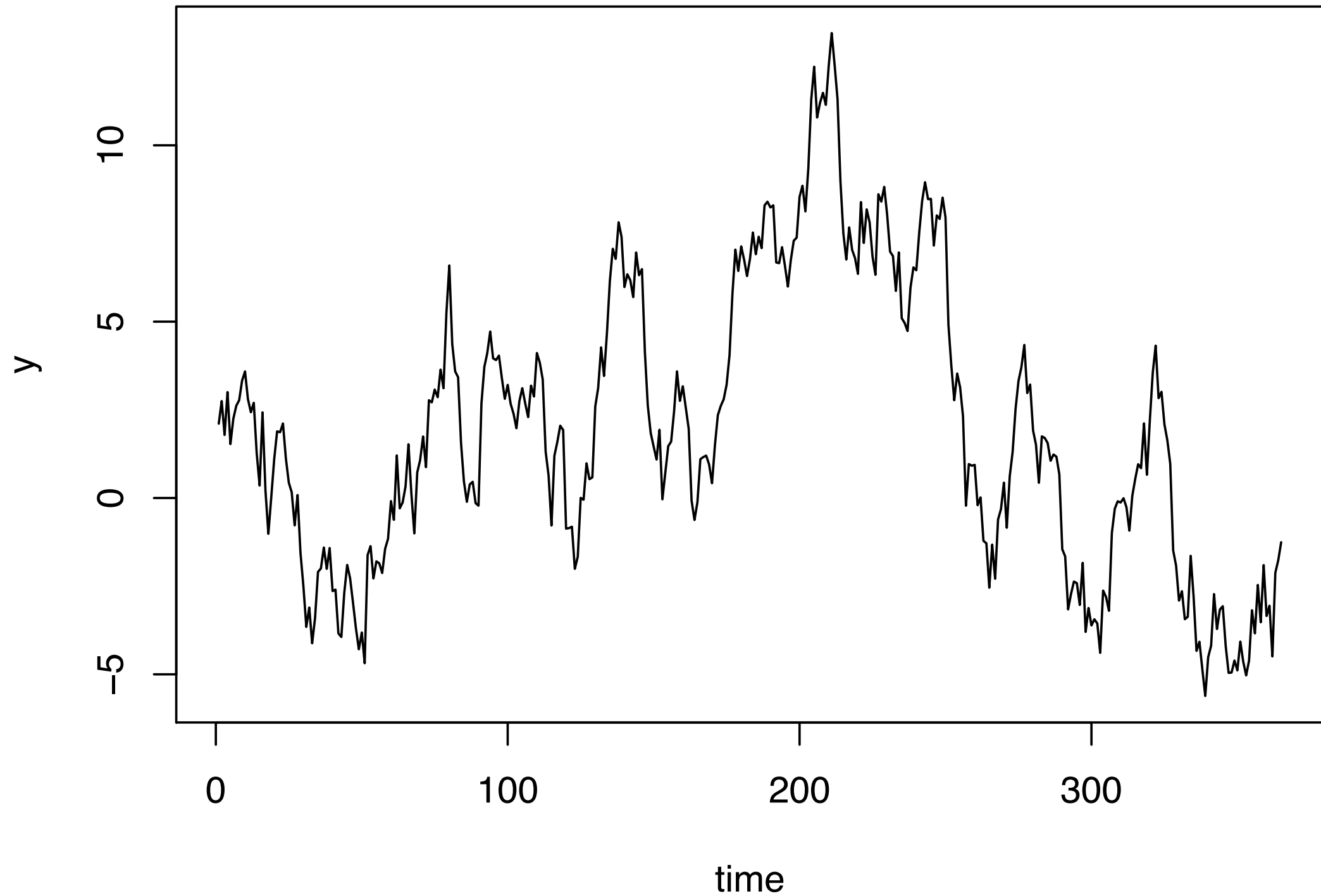
# Random vs. Fixed Variation

- Like with any statistical modeling it is important to separate out what is **random** and what is **fixed**
- Traditional time series modeling methods focus on modeling the random aspects (often no mention of fixed aspects)
- In environmental health applications, often many things are fixed
  - season, day-of-week, temperature, etc.
  - identifying those fixed effects is key part of science
- Residual variation may still be autocorrelated
- Because you often only observe a single time series, it is not possible for a model to determine what is fixed and what is random

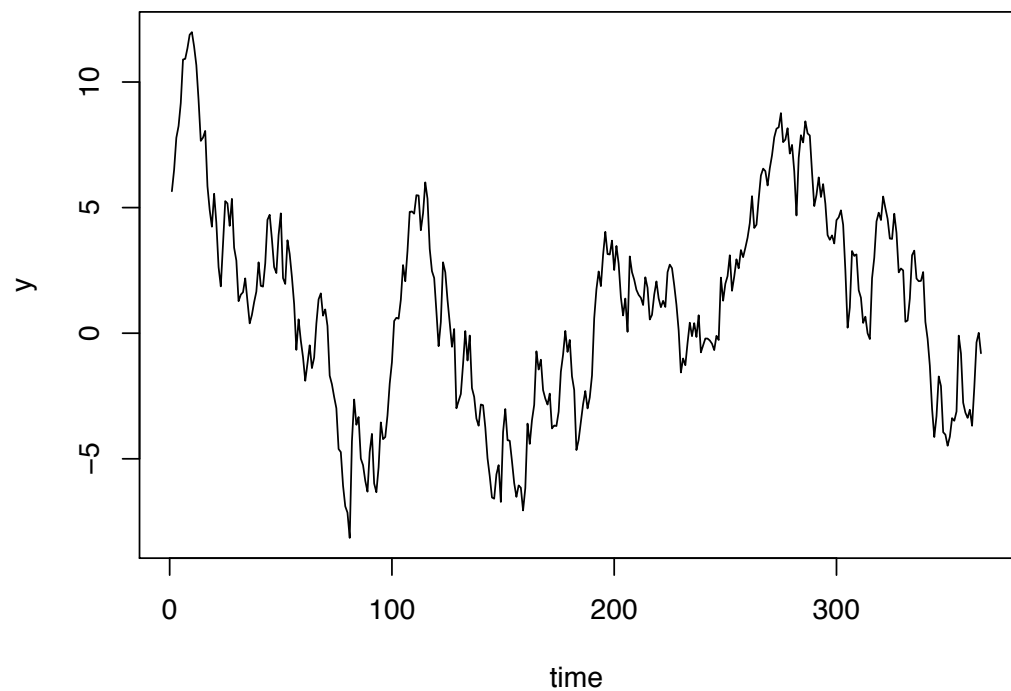
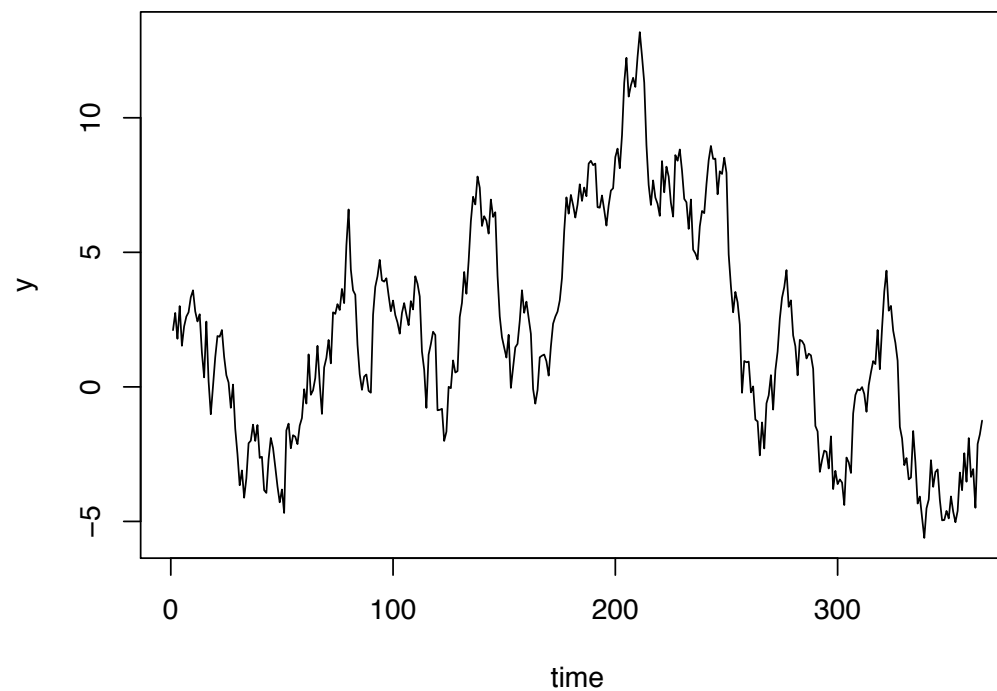
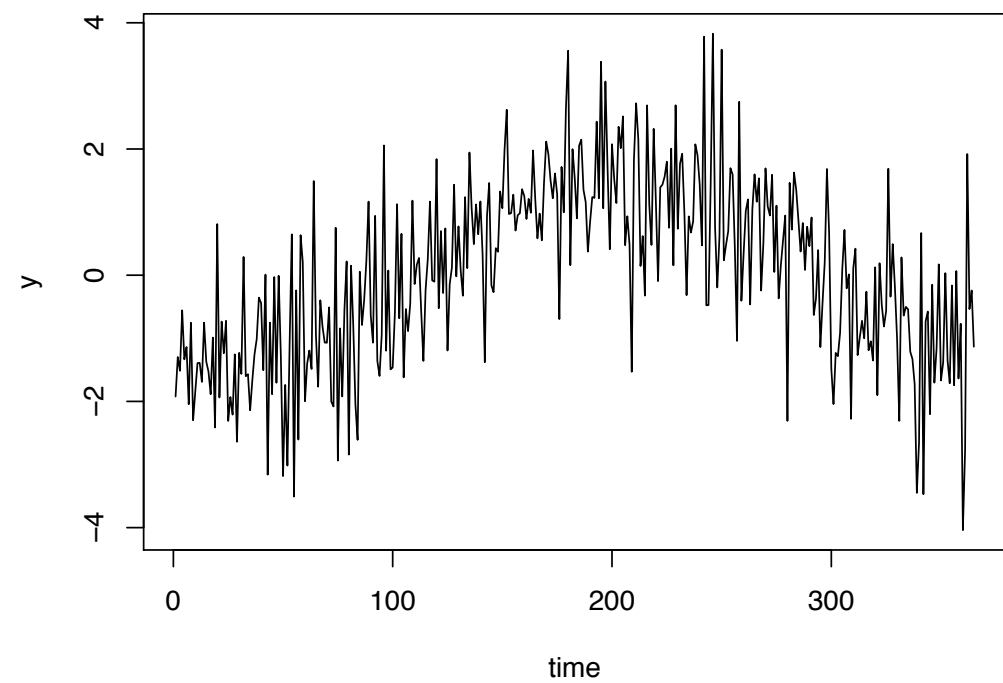
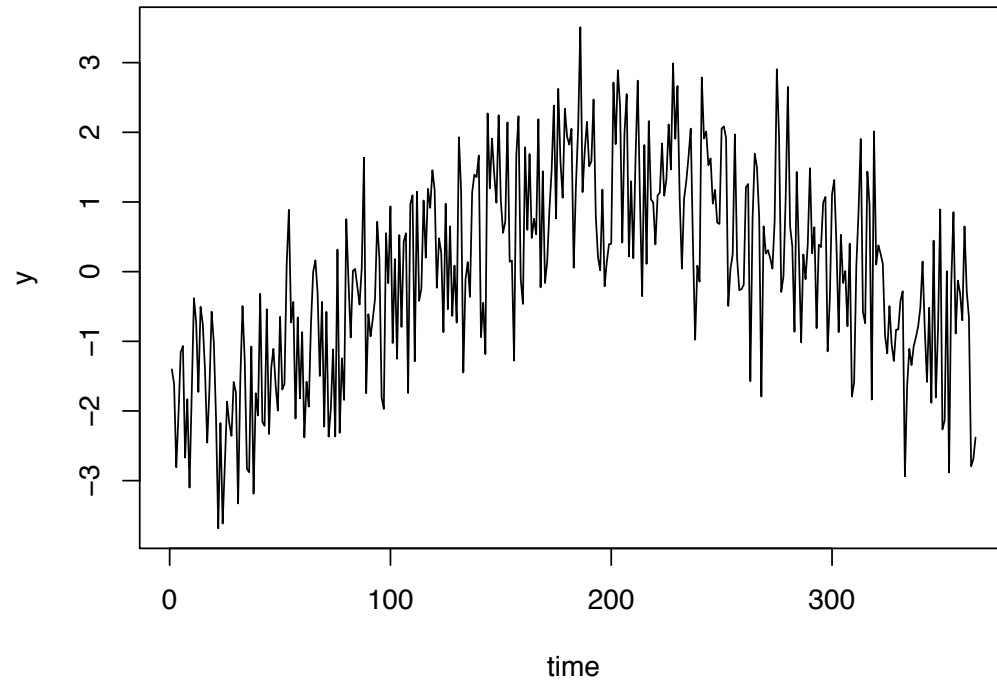
# Random or Fixed?



# Random or Fixed?



# Random or Fixed?



# Time Series Analysis in Environmental Health

- Primary task:
  - Model  $Y$  vs.  $X$  controlling for  $Z$
  - We usually do *not* want to predict  $Y$  from a set of predictors  $X$
  - We usually do *not* want to predict a future value of  $Y$  from a set of predictors  $X$
- Key problems:
  - Temporal misalignment
  - Missing data is often a problem
- Most concerned with *residual* autocorrelation after removing fixed effects

# Example: pDRs and COPD

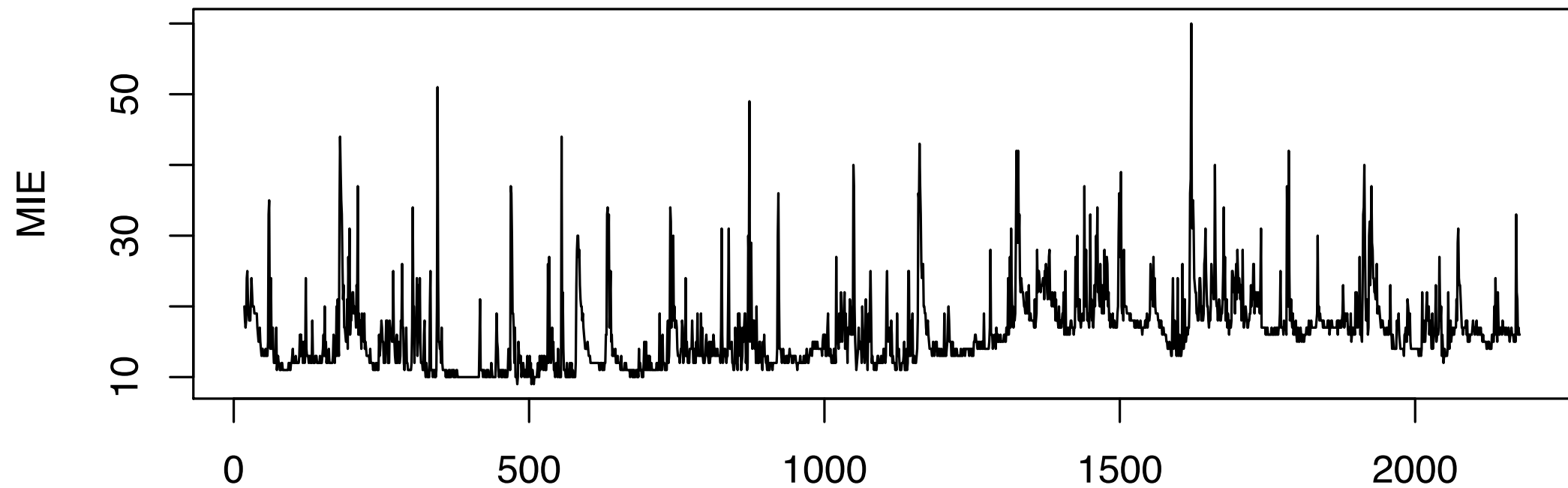
- A study of elderly adults with COPD living in Baltimore, MD; total of 85 subjects
- Longitudinal followup with 1 visit every 3 months (total 3 visits)
- Personal DataRAM (pDR) placed in home to measure PM<sub>2.5</sub> at 5-minute intervals over 7-days
- Are indoor levels of PM<sub>2.5</sub> associated with COPD?



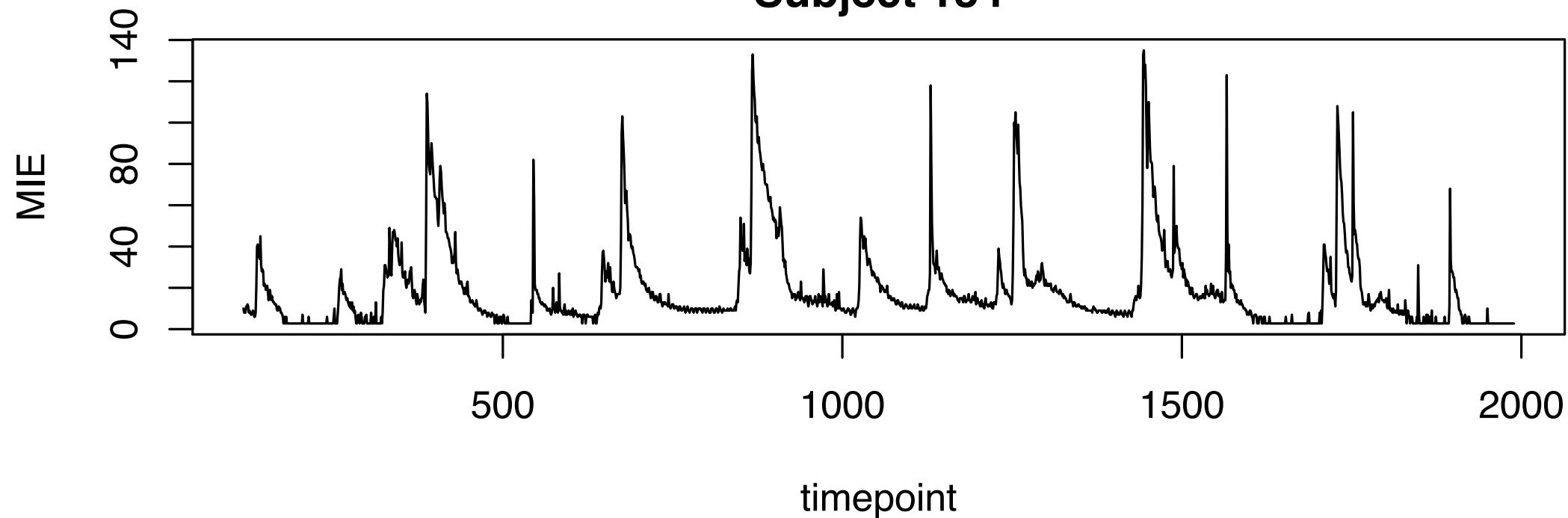


# Real-Time PM<sub>2.5</sub>

**Subject 136**

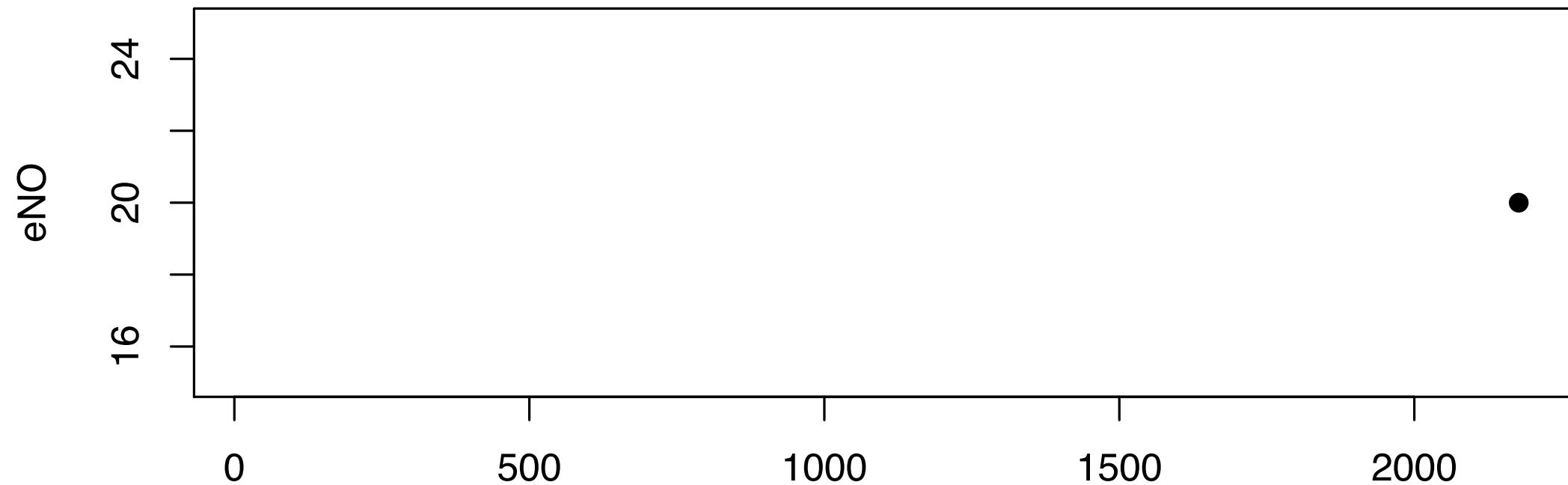


**Subject 134**

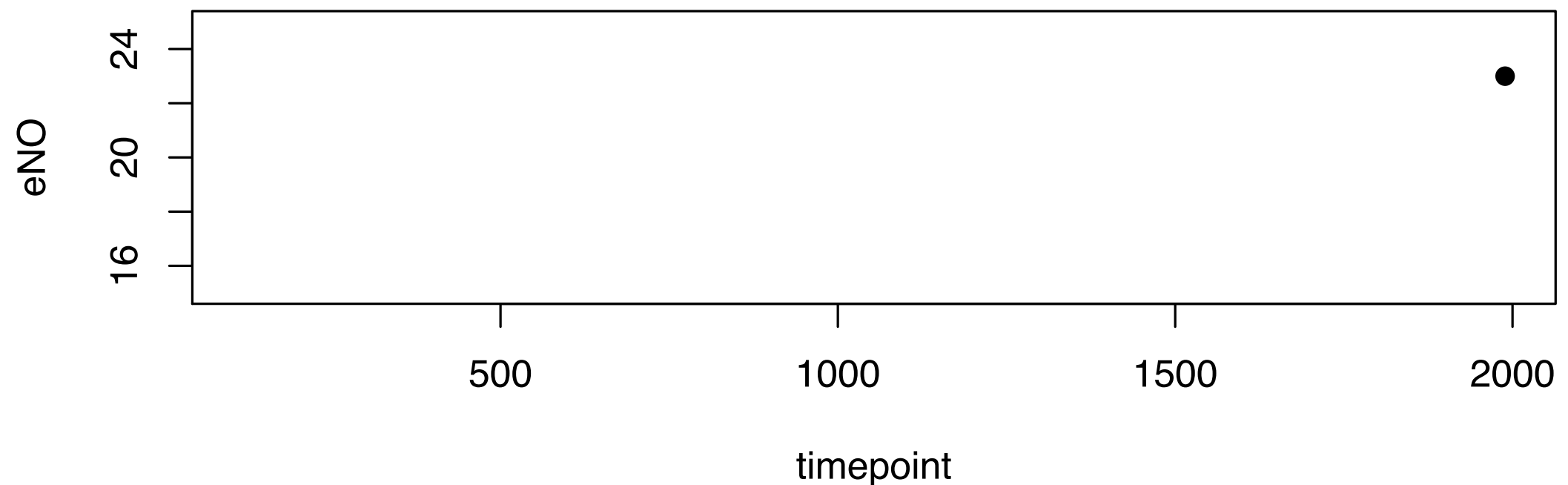


# Less-Than-Real-Time Outcome

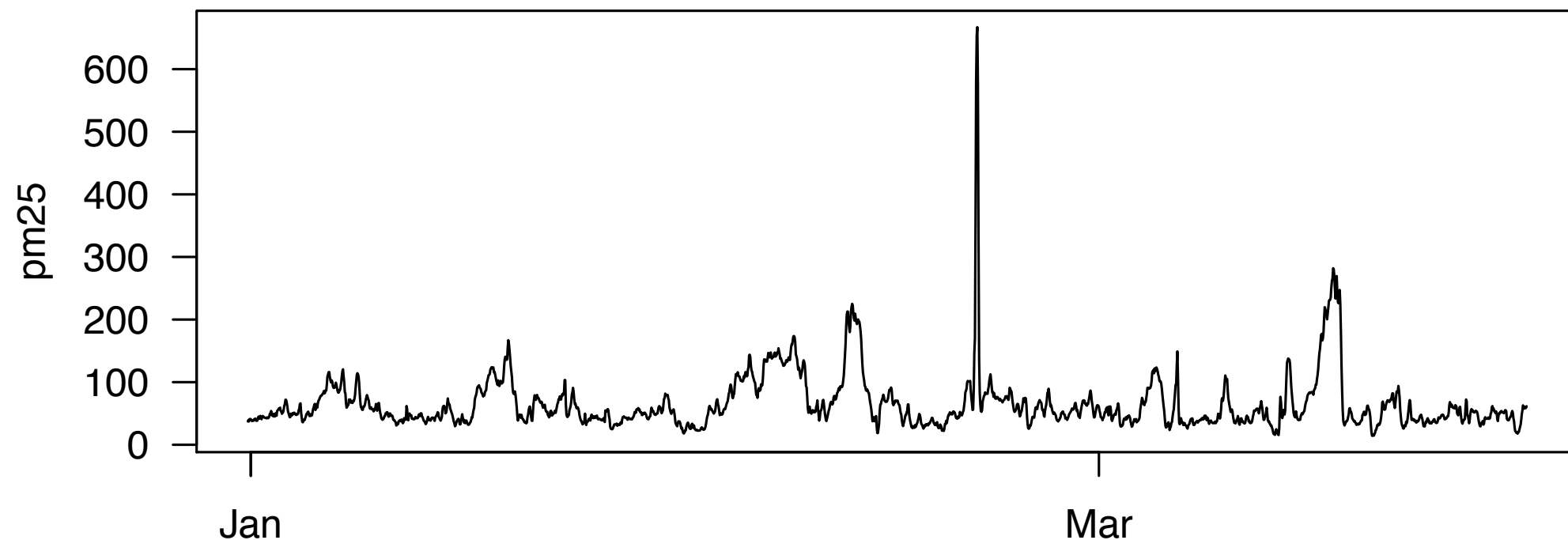
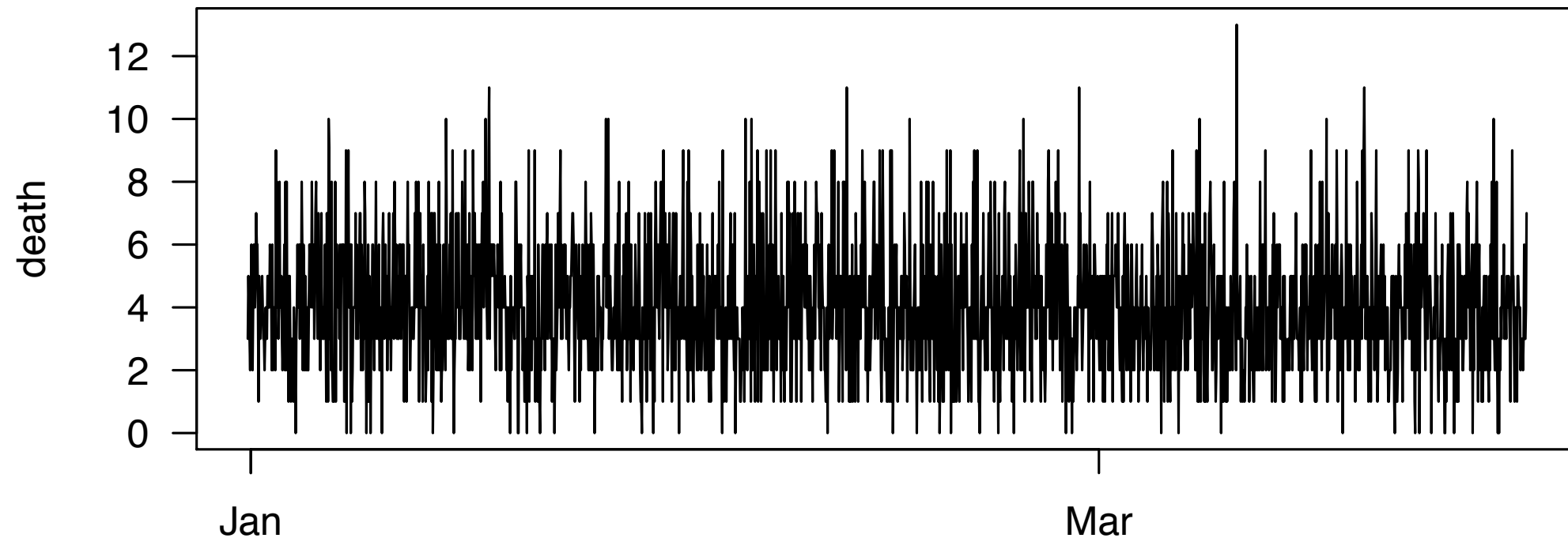
**Subject 136**



**Subject 134**



# Hourly PM2.5 and Mortality (Seoul, Korea)



# Estimation or Prediction?

- One approach is to design models to optimally **predict the outcome**
- Another approach is to develop models that **estimate an association** and adequately control for confounding
- These two approaches do not necessarily lead to the same model!
- If X and Y have an inherently weak relationship, then it makes little sense to include X in a model for optimally predicting Y
  - e.g. step-wise model selection methods will usually remove X from the model; then what?
  - Potential confounders weakly correlated with Y (but perhaps strongly correlated with X) may not be included in a prediction model

# Confounding in Time Series

- A confounder is something that is associated with both the outcome and the risk factor
- In time series studies, the association between a risk factor and outcome is potentially confounded by things that vary in time (day to day, week to week)
  - e.g. weather, season, temperature, pollutants, long-term trends
- Things that “do not vary over time” are not confounders

# Timescale of Variation

- In time series analysis the *timescale of variation* is an important aspect to consider
- Do we care about year-to-year, month-to-month, or day-to-day variation in X and Y?
- On what timescales do potential confounders vary?
- If a potential confounder does not vary on your timescale of interest, then it is not a confounder.

# Timescale Decomposition

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$


$$x_t = \bar{x}_t^Y + (x_t - \bar{x}_t^Y)$$


$$y_t = \beta_0 + \beta_1 \bar{x}_t^Y + \beta_2 (x_t - \bar{x}_t^Y) + \varepsilon_t$$


If  $\beta_1 = \beta_2$ , then we have the first model

# Timescale Decomposition

$$y_t = \beta_0 + \beta_1 \bar{x}_t^Y + \beta_2 (x_t - \bar{x}_t^Y) + \varepsilon_t$$


$$z_t = \bar{z}_t^S + (z_t - \bar{z}_t^S)$$


$$y_t = \beta_0 + \beta_1 \bar{x}_t^Y + \beta_2 \bar{z}_t^S + \beta_3 (z_t - \bar{z}_t^S) + \varepsilon_t$$


$$u_t = \bar{u}_t^W + (u_t - \bar{u}_t^W)$$



# Timescale Decomposition

$$u_t = \bar{u}_t^W + (u_t - \bar{u}_t^W)$$

$$x_t = \bar{x}_t^Y + \bar{z}_t^S + \bar{u}_t^W + r_t$$

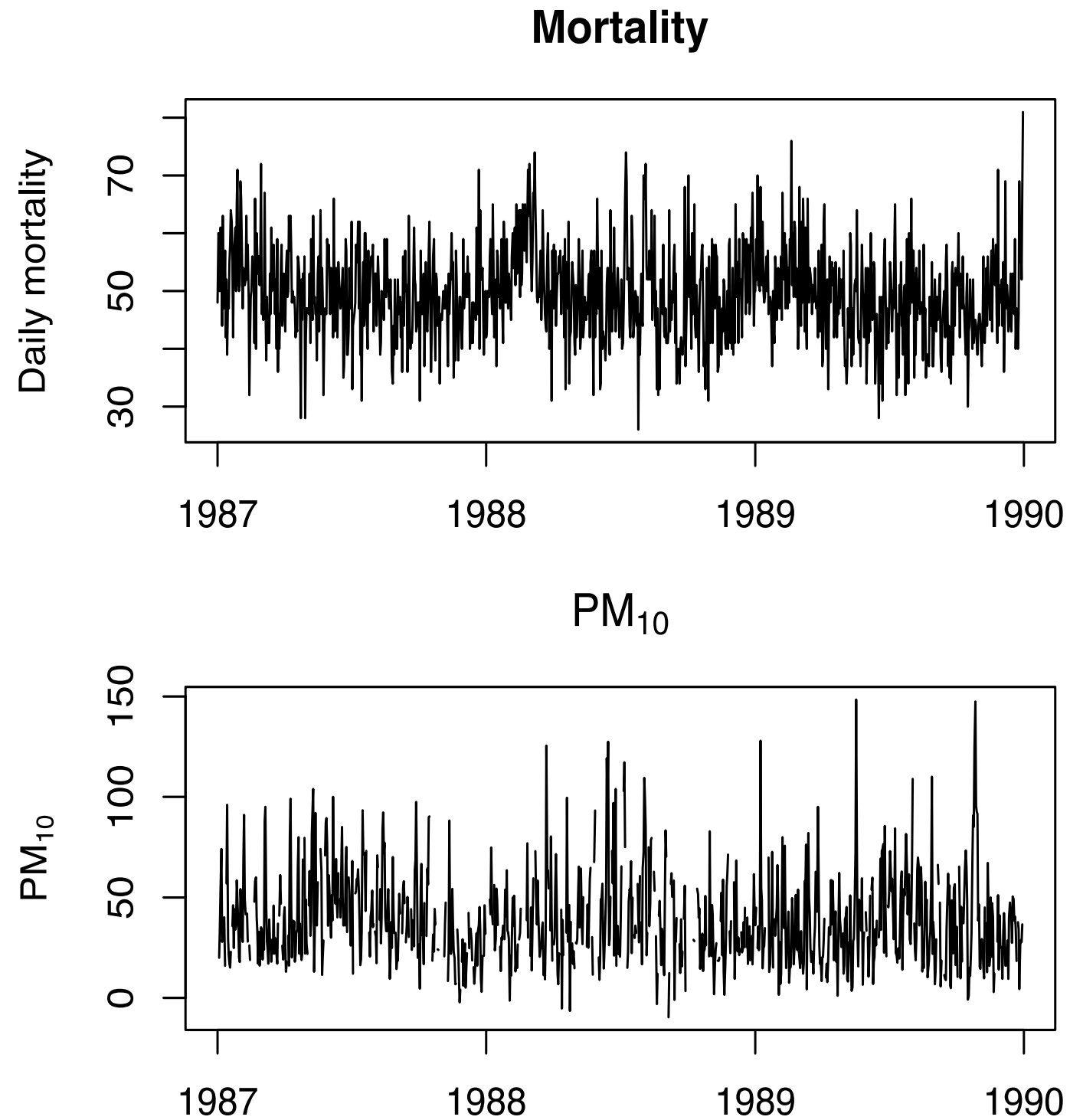
$$y_t = \beta_0 + \beta_1 \bar{x}_t^Y + \beta_2 \bar{z}_t^S + \beta_3 \bar{u}_t^W + \beta_4 r_t + \varepsilon_t$$

365-day MA

90-day MA

7-day MA

# Detroit Mortality and PM<sub>10</sub>

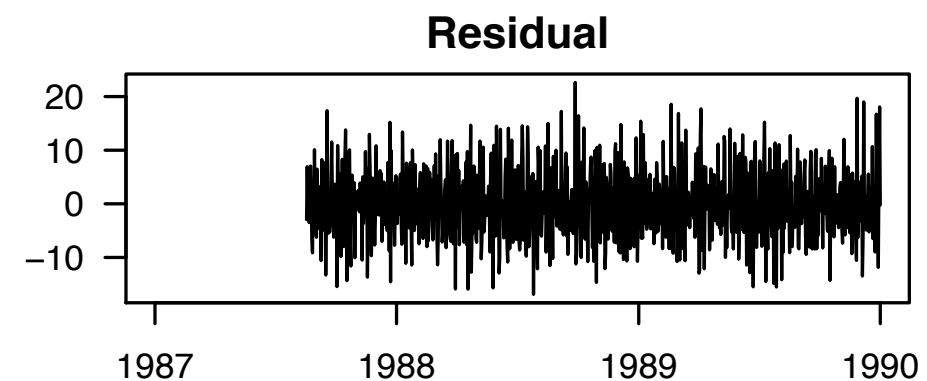
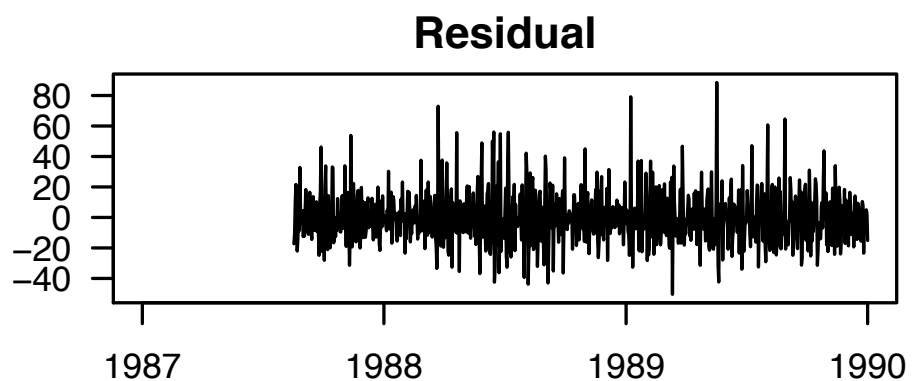
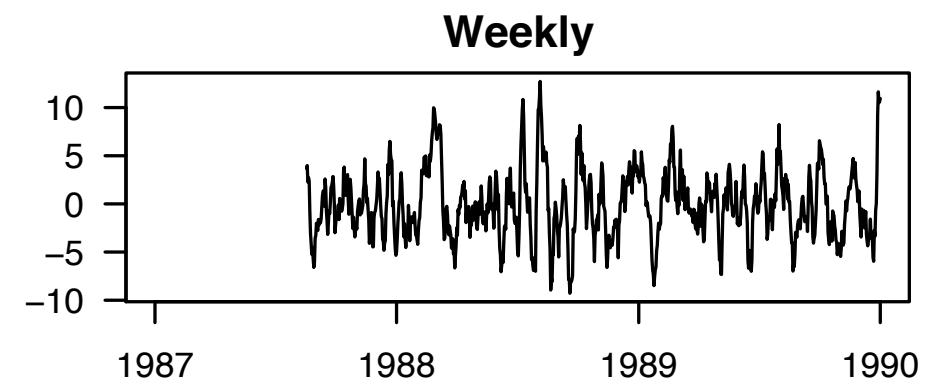
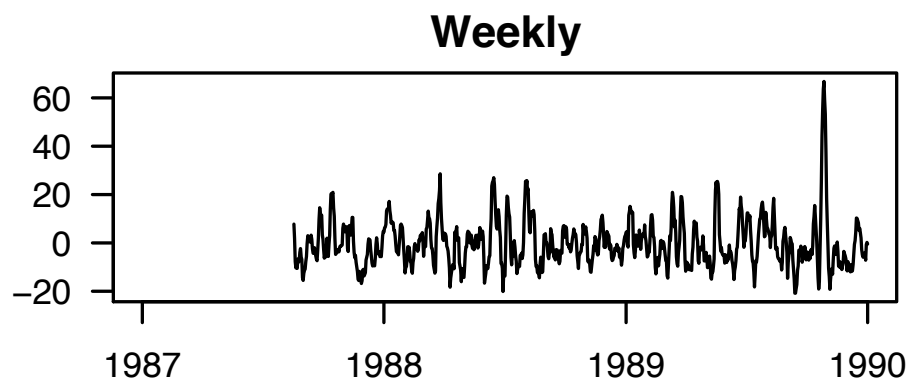
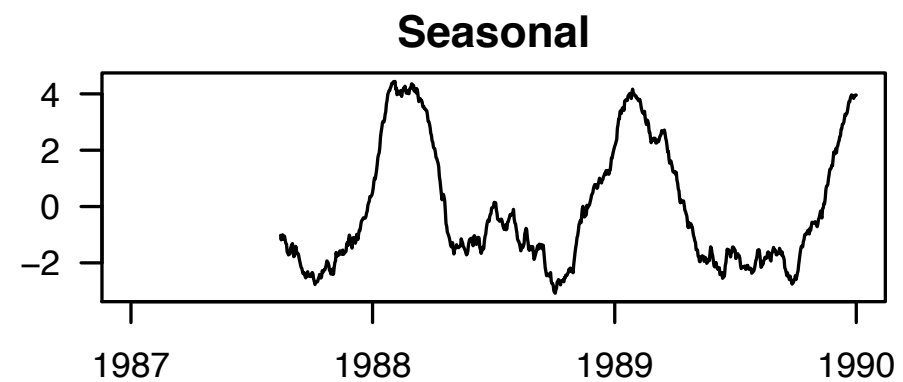
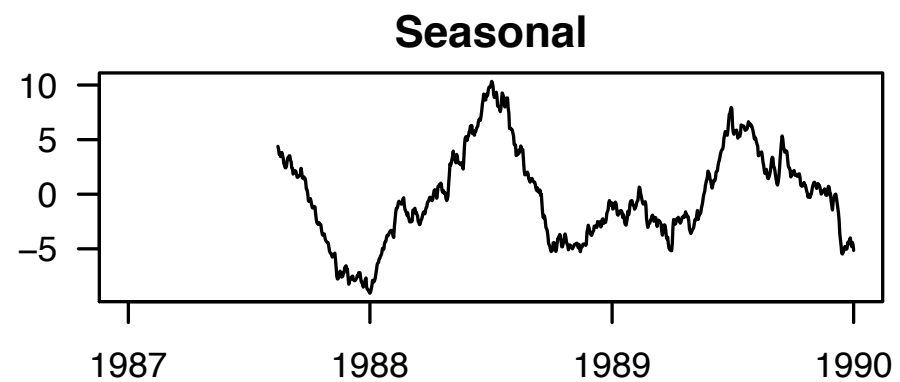
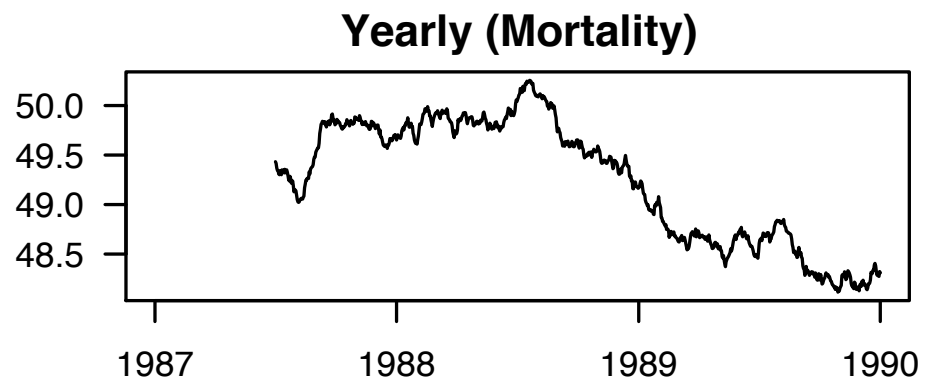
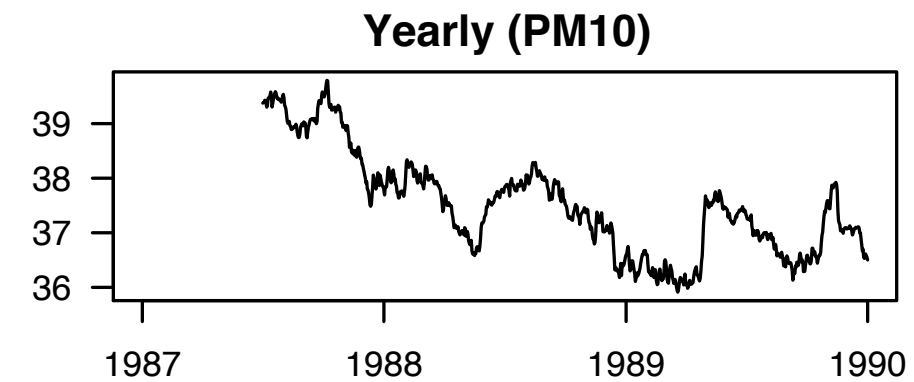


# Detroit PM<sub>10</sub> and Mortality

$$Y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	46.1798	0.2263	204.11	0.0000
x	0.0232	0.0057	4.06	0.0000

# Timescales of Variation



# Detroit PM<sub>10</sub> and Mortality

$$Y_t = \beta_0 + \beta_1 \mathbf{yearly}_t + \beta_2 \mathbf{seasonal}_t + \beta_3 \mathbf{weekly}_t + \beta_4 \mathbf{residual}_t + \varepsilon_t$$

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	34.1031	1.3098	26.04	0.0000
x.yearly	0.3783	0.0383	9.88	0.0000
z.seasonal	-0.4354	0.0295	-14.76	0.0000
u.weekly	0.0532	0.0123	4.33	0.0000
r	0.0215	0.0070	3.07	0.0022

# Which Estimate Do We Use?

- Which timescale estimate is most appropriate or useful (i.e. tells the best or interesting “story”)?
- Which potential confounders affect which timescale estimate and how?
- At which timescale do we have the best data, measurements, knowledge/hypotheses?

# Regression Model Assumptions

- Errors are independent of covariates
- Errors have mean zero
- Errors are independent of each other (autocorrelation?)
- Variance of errors is constant and  $> 0$
- [Errors are Normally distributed with mean 0 and variance  $\sigma^2$ ]

# Time Series Regression

## Model Assumptions

- In a time series regression model, errors may be correlated
  - Why are errors correlated? Can we explain it?
- Correlated errors (if ignored) can affect estimation of parameters
  - Is that important? Do we need to model it?
- What is random? What is fixed?



# Detroit PM<sub>10</sub> and Mortality Analysis

Call:

lm(formula = y ~ x)

$$Y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$

Residuals:

Min	1Q	Median	3Q	Max
-0.83914	-0.11227	0.00688	0.11807	0.55958

Coefficients:

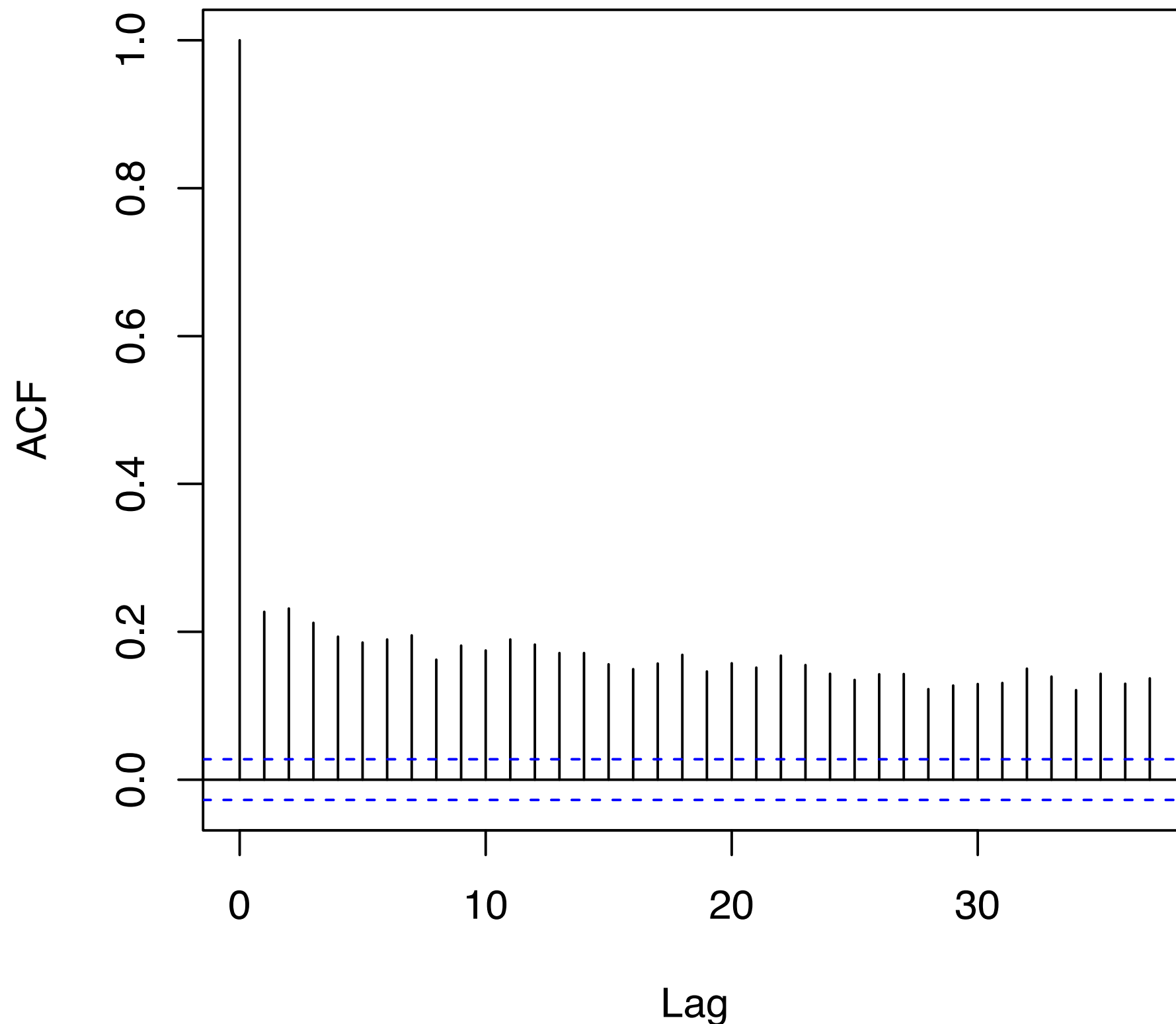
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	3.8348985	0.0024221	1583.311	<2e-16
x	0.0003014	0.0001253	2.406	0.0162

Residual standard error: 0.1732 on 5112 degrees of freedom

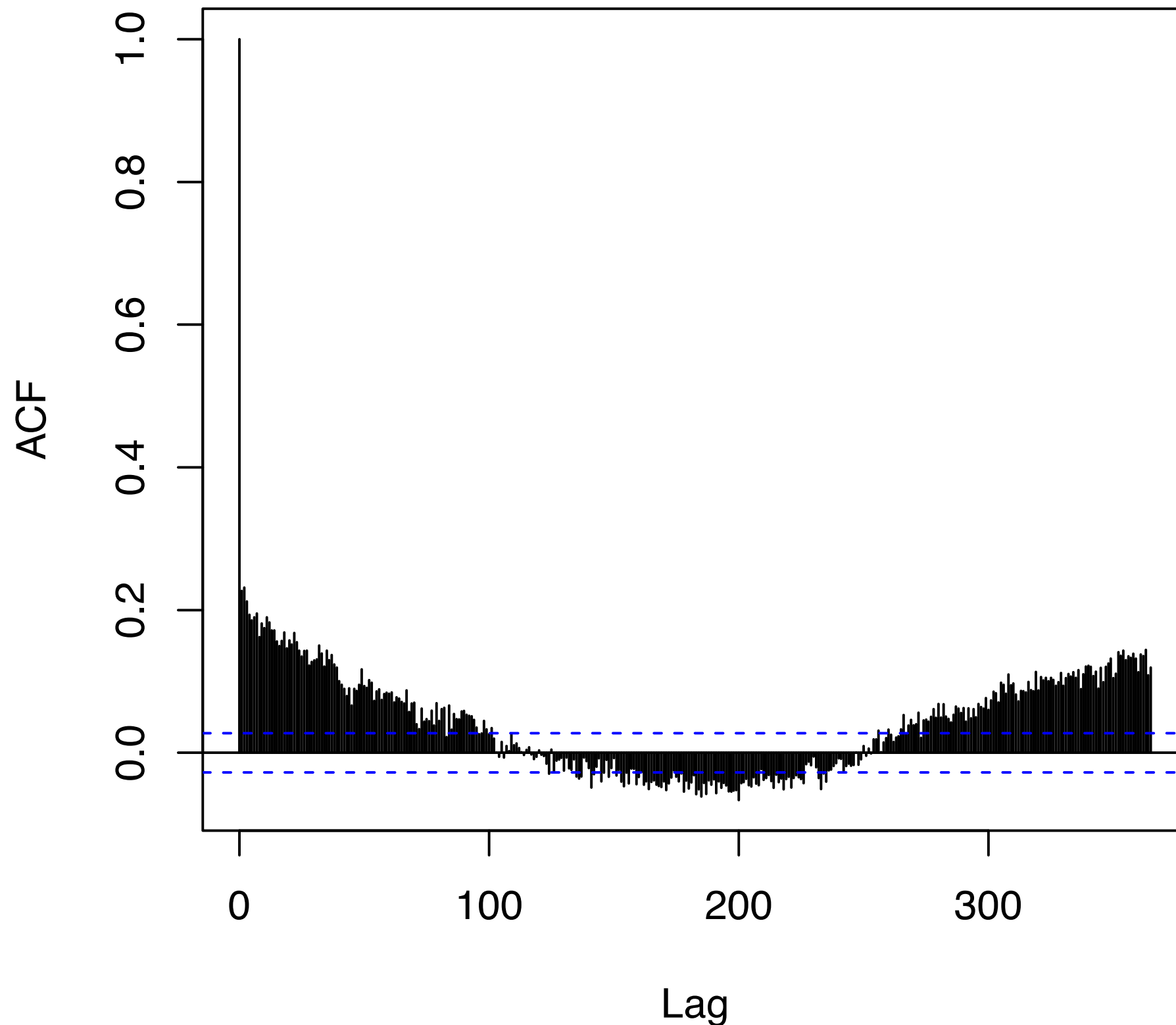
Multiple R-squared: 0.001131,      Adjusted R-squared:  
0.0009357

F-statistic: 5.788 on 1 and 5112 DF,    p-value: 0.01617

# Residual autocorrelation?



# Residual autocorrelation?



# Removing season

Call:

```
lm(formula = y ~ season + x)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.81887	-0.10184	0.00779	0.11293	0.55844

Coefficients:

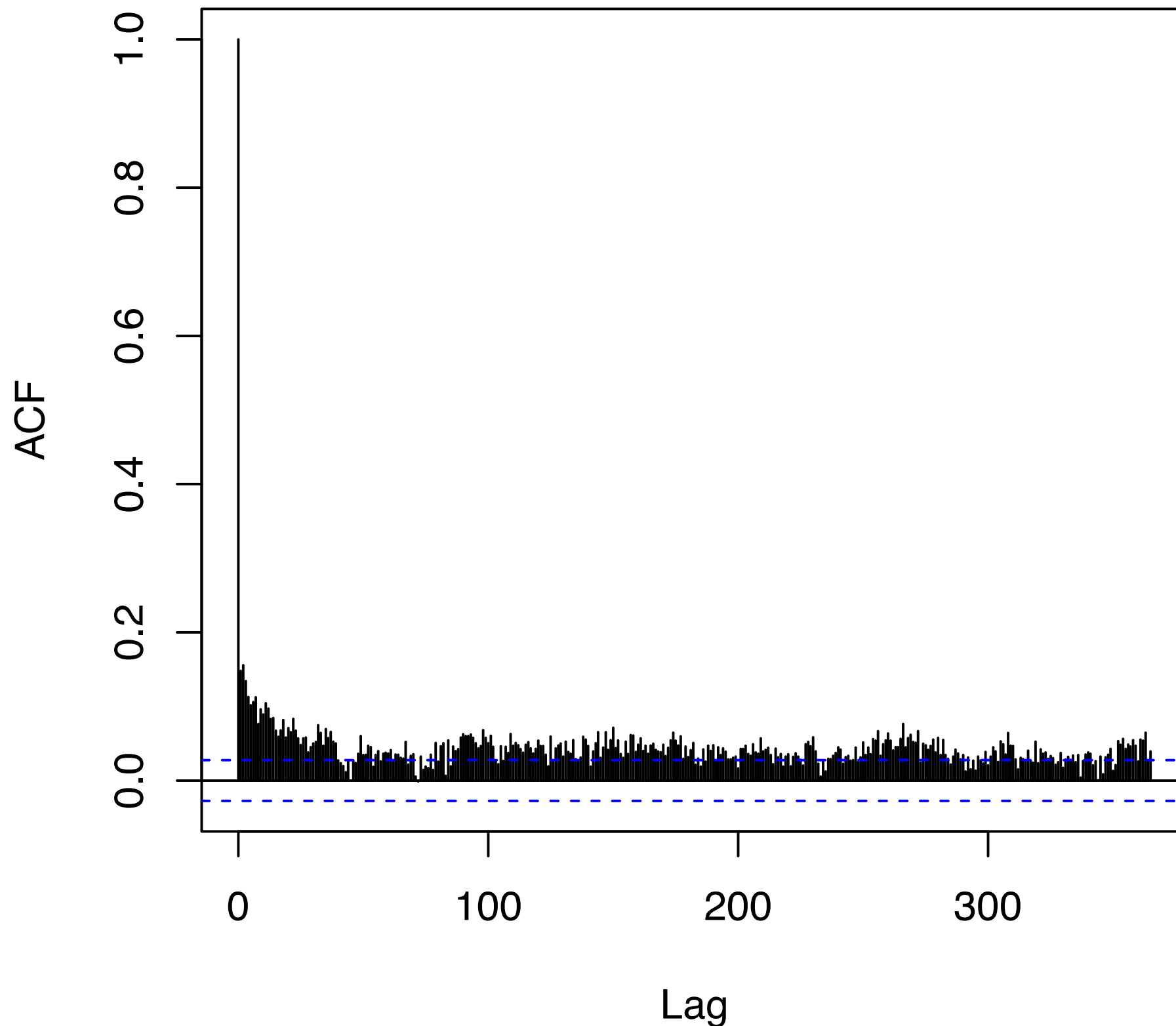
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	3.9155533	0.0046581	840.587	< 2e-16
seasonQ2	-0.1008966	0.0065837	-15.325	< 2e-16
seasonQ3	-0.1448564	0.0065731	-22.038	< 2e-16
seasonQ4	-0.0754779	0.0065410	-11.539	< 2e-16
x	0.0005982	0.0001205	4.965	0.000000711

Residual standard error: 0.1652 on 5109 degrees of freedom

Multiple R-squared: 0.09179, Adjusted R-squared: 0.09108

F-statistic: 129.1 on 4 and 5109 DF, p-value: < 2.2e-16

# Residual autocorrelation?



# Remove season + trend

Call:

```
lm(formula = y ~ season + day + x)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.76282	-0.10167	0.00566	0.10936	0.55349

Coefficients:

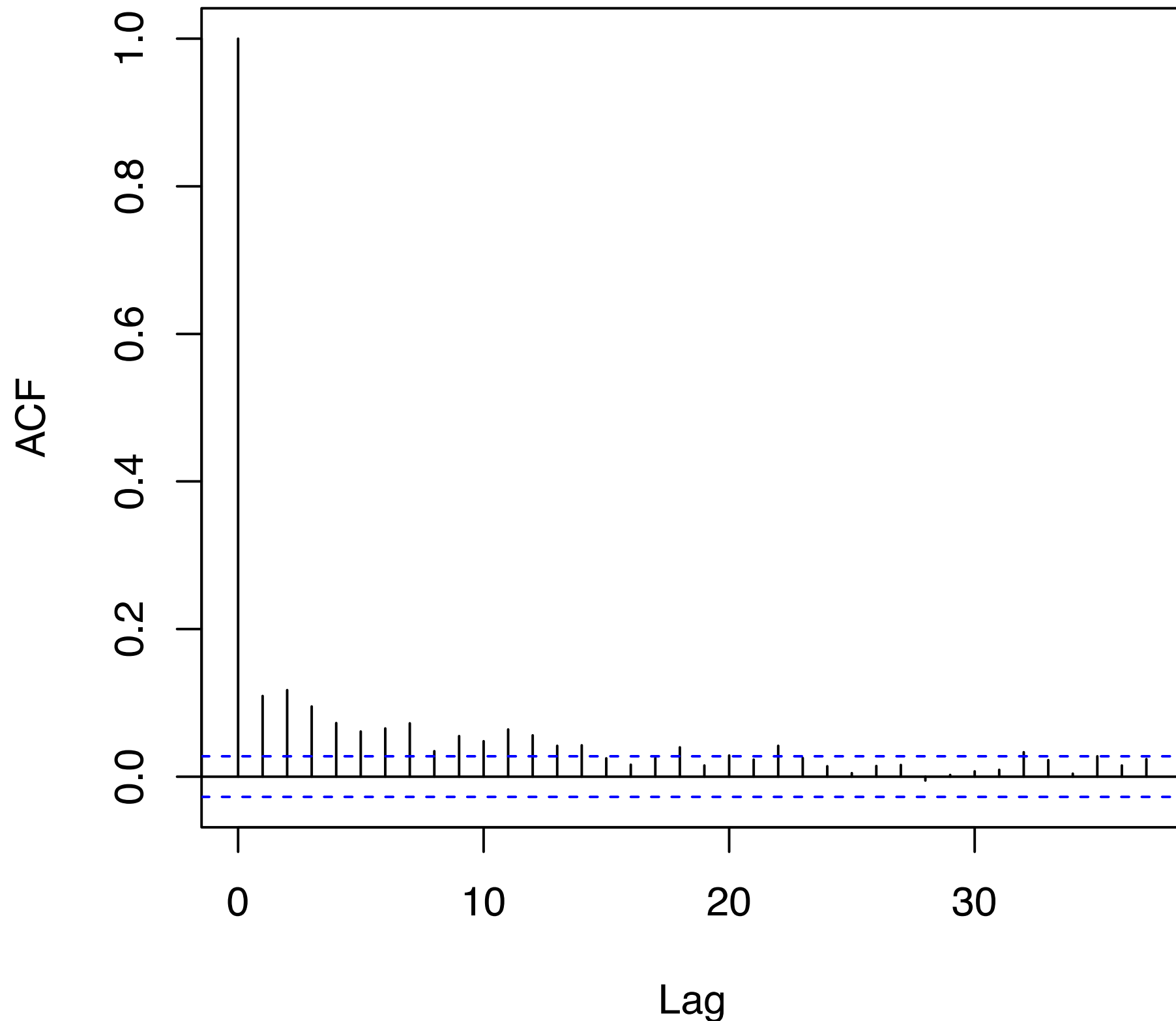
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	4.117496865	0.013998439	294.140	< 2e-16
seasonQ2	-0.098744775	0.006440817	-15.331	< 2e-16
seasonQ3	-0.140557052	0.006435061	-21.842	< 2e-16
seasonQ4	-0.069072422	0.006411230	-10.774	< 2e-16
day	-0.000023407	0.000001534	-15.257	< 2e-16
x	0.000589037	0.000117849	4.998	0.000000598

Residual standard error: 0.1616 on 5108 degrees of freedom

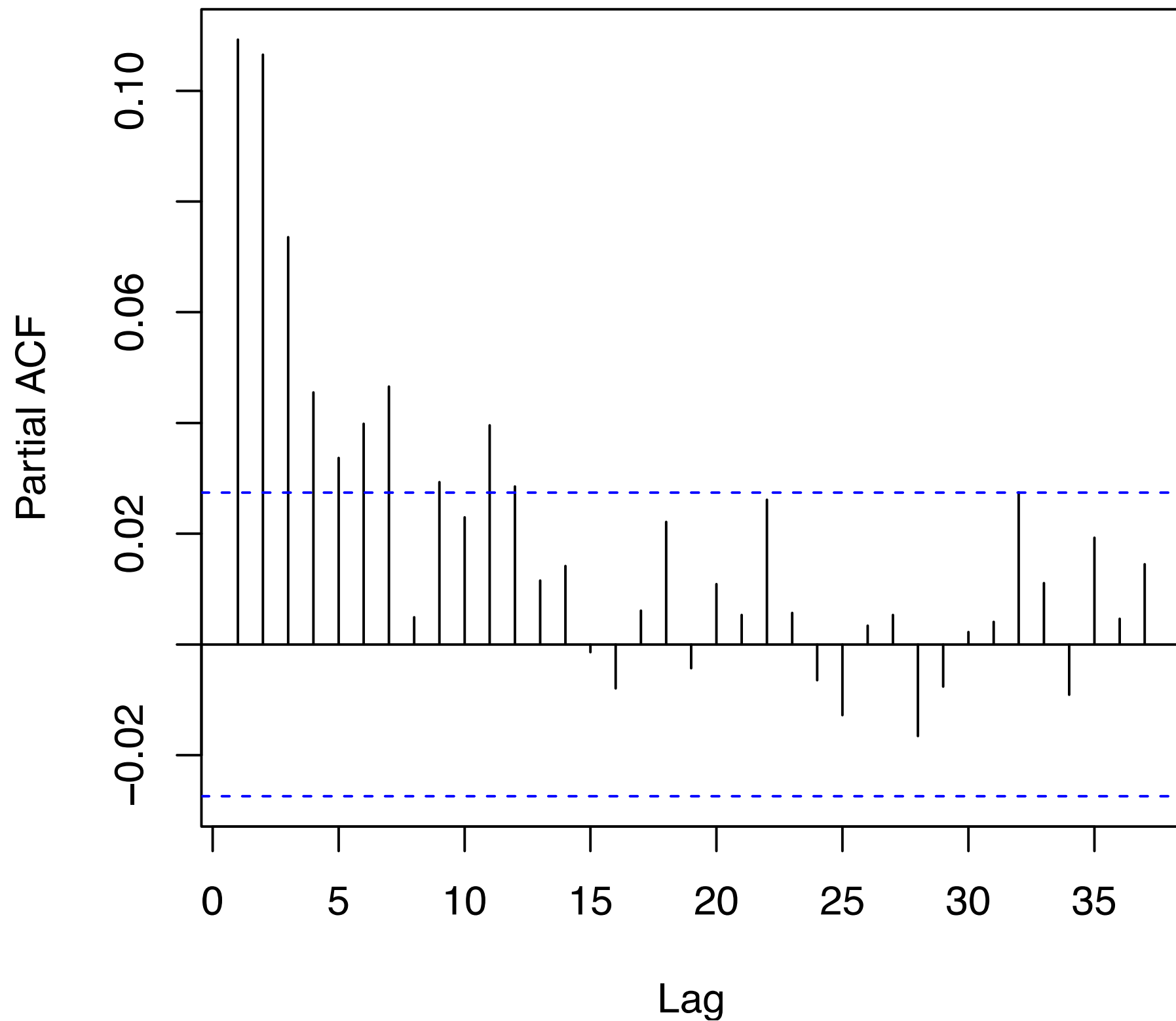
Multiple R-squared: 0.1314, Adjusted R-squared: 0.1305

F-statistic: 154.5 on 5 and 5108 DF, p-value: < 2.2e-16

# Residual autocorrelation?



# PACF Plot





# AR(6) model results

Call:

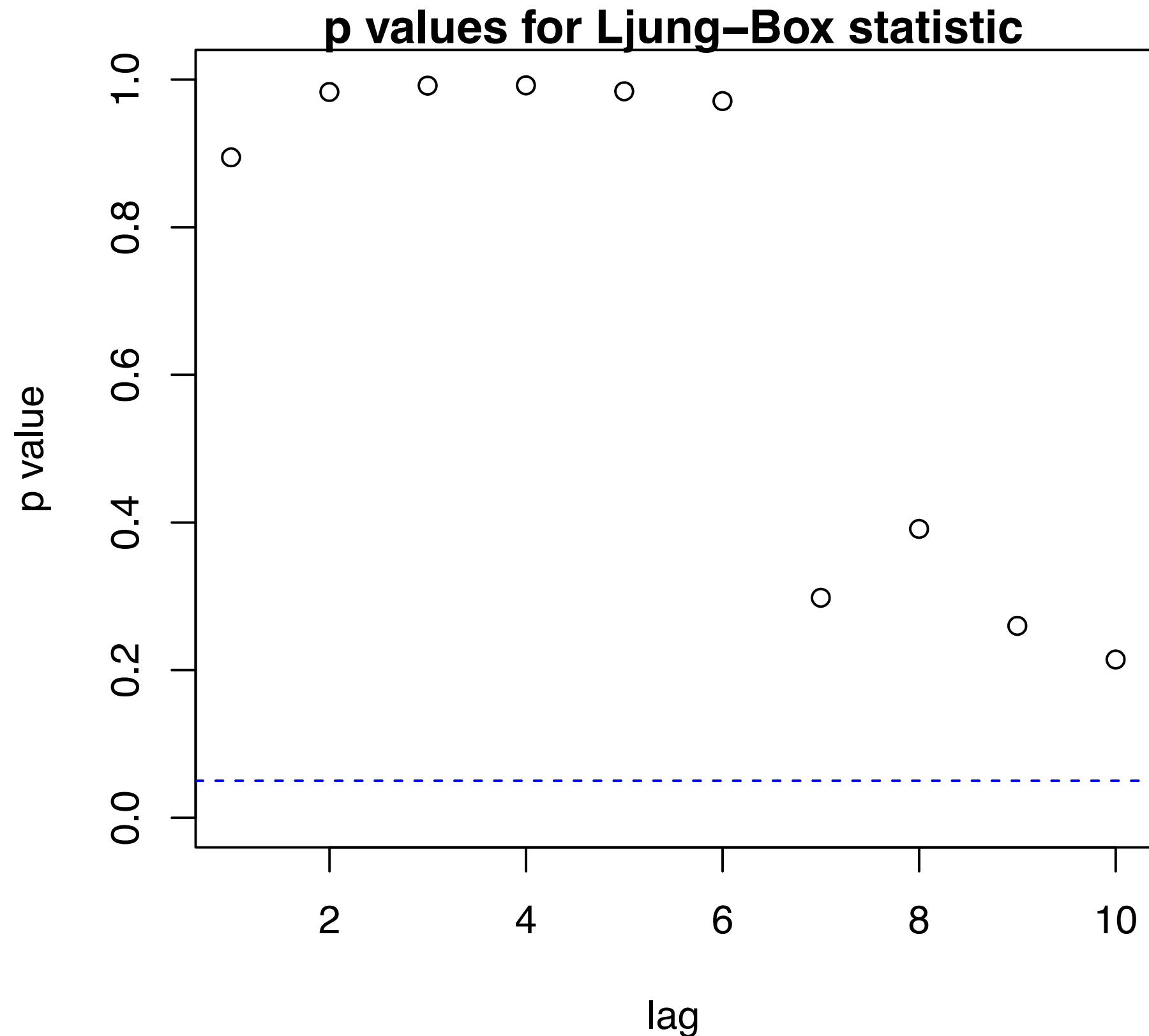
```
arima(x = y, order = c(6, 0, 0), xreg = mm, include.mean = F)
```

Coefficients:

	ar1	ar2	ar3	ar4	ar5	ar6	(Intercept)	seasonQ2
	0.0834	0.0909	0.0639	0.0392	0.0307	0.0403	4.1137	-0.0938
s.e.	0.0140	0.0140	0.0141	0.0141	0.0140	0.0140	0.0211	0.0094
	seasonQ3	seasonQ4	day	x				
	-0.1356	-0.0670	0	0.0006				
s.e.	0.0095	0.0094	0	0.0001				

sigma^2 estimated as 0.02521: log likelihood = 2154.22, aic = -4282.44

# Test for residual autocorrelation



# Robust variance or AR model?

	Coef	Naive	Robust	AR(6)
(Intercept)	4.117497	0.013998	0.017070	0.020225
seasonQ2	-0.098745	0.006441	0.007945	0.009088
seasonQ3	-0.140557	0.006435	0.008498	0.009153
seasonQ4	-0.069072	0.006411	0.007999	0.009058
day	-0.000023	0.000002	0.000002	0.000014
x	0.000589	0.000118	0.000131	0.000120

# Summary

- Time series data relate changes over time of an exposure and outcome
- Different models can be used for estimation/explanation and prediction
- Understanding timescales of variation for exposure and outcome is important in time series analysis
- Autocorrelation in residuals can indicate unexplained variability (but often can be explained!)

# Literature

