Autocorrelation 1: Temporal Autocorrelation

Michael Noonan

Biol 520C: Statistical modelling for biological data

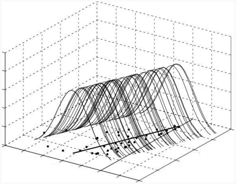
Table of contents

1. Autocorrelation and the IID Assumption
2. Temporal Autocorrelation
3. Correcting Temporal Autocorrelation
4. Correcting Temporal Autocorrelation in `R`

Autocorrelation and the IID Assumption

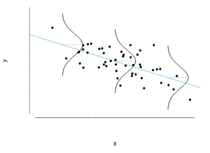
The IID Assumption

Last lecture we saw how differences in variances across groups (i.e., heteroskedasticity) can break the 'identical' part of the IID assumption.



We also saw how adding a variance structure to the model can correct for heteroskedasticity.

Over the next few lectures we'll explore ways in which the 'independent' part of the IID assumption can be broken, what the implications of this mean, and how to correct for it.



Autocorrelation is important

© Chris Sorensen

Dr. Sam Wang, Neuroscientist
—Princeton Election Consortium

© ABC News

Nate Silver, Statistician
—FiveThirtyEight.com

Autocorrelation is important

© Chris Sorensen

Dr. Sam Wang, Neuroscientist
—Princeton Election Consortium

© ABC News

Nate Silver, Statistician
—FiveThirtyEight.com

"It is totally over. If Trump wins more than 240 electoral votes, I will eat a bug."

Autocorrelation is important

© Chris Sorensen

Dr. Sam Wang, Neuroscientist
—Princeton Election Consortium

"It is totally over. If Trump wins more than 240 electoral votes, I will eat a bug."

© ABC News

Nate Silver, Statistician
—FiveThirtyEight.com

"Trump Is Just A Normal Polling Error Behind Clinton."

Autocorrelation is important

© Chris Sorensen

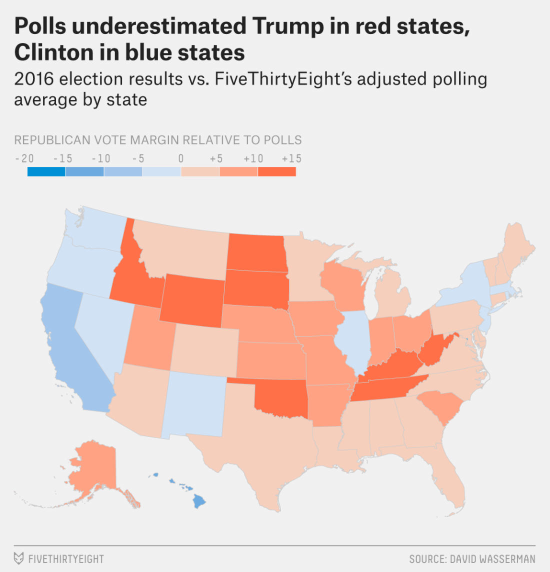
© ABC News

Nate Silver, Statistician

—FiveThirtyEight.com

“Trump Is Just A Normal Polling Error Behind Clinton”

Ignoring non-ind. → overconfidence



Polls underestimated Trump in red states, Clinton in blue states

2016 election results vs. FiveThirtyEight's adjusted polling average by state

SOURCE: DAVID WASSERMAN

Are these polling errors independently distributed?

This same statistical issue that caused overly confident predictions of Clinton's 2016 victory can result in overconfidence in parameter estimates and predictions in regression models.

Autocorrelation impact

Sample size, n is the denominator when calculating both SEs and CIs.

$$\mathrm{SE} = \tfrac{\sigma}{\sqrt{n}} \qquad\qquad 95\%\mathrm{CI} = \bar{x} \pm 1.96\tfrac{\sigma}{\sqrt{n}}$$

All else equal: $\uparrow n = \downarrow \mathrm{SE}\ \&\ \downarrow \mathrm{CI}$

But with autocorrelated data each new datapoint is related to a previously collected datapoint and does not bring a full independent datapoint worth of information (e.g., 90% autocorr. $\approx$ 10% new info).

When data are autocorrelated $n_{\mathrm{effective}} < n$, meaning SEs and CIs shrink faster than they should, resulting in a false sense of confidence.

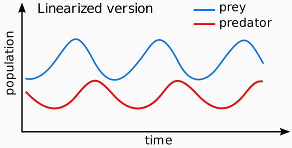
Effect is usually strongest on SEs and CIs, but autocorrelation can also impact the mean: $\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \frac{x_1 + \cdots + x_n}{n}$

Sources of autocorrelation

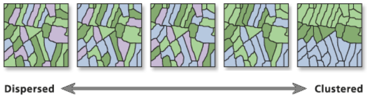
Anything that causes some data points to be more similar to each other than others can result in autocorrelation.

Over the next three lectures we will be covering the three most common sources of autocorrelation in biological data:

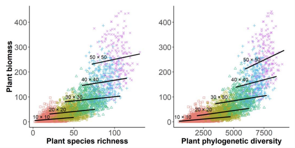
- **Time**: Data that are close together in time are more related.



- **Space**: Data that are close together in space are more related.



- **Phylogeny**: Species that are closer together on an evolutionary timescale are more related.



(Liang *et al.*, 2019)

Temporal Autocorrelation

Temporal Autocorrelation

As biologists, we often find ourselves measuring things over time (e.g., population sizes, animal locations, CO_2 levels, reaction times, etc.).

When this is the case, autocorrelation can arise in two ways:

1. The autocorrelation timescale of the process we're measuring.

Temporal Autocorrelation

As biologists, we often find ourselves measuring things over time (e.g., population sizes, animal locations, CO_2 levels, reaction times, etc.).

When this is the case, autocorrelation can arise in two ways:

1. The autocorrelation timescale of the process we're measuring.

Temporal Autocorrelation

As biologists, we often find ourselves measuring things over time (e.g., population sizes, animal locations, CO_2 levels, reaction times, etc.).

When this is the case, autocorrelation can arise in two ways:

1. The autocorrelation timescale of the process we're measuring.

Temporal Autocorrelation

As biologists, we often find ourselves measuring things over time (e.g., population sizes, animal locations, CO_2 levels, reaction times, etc.).

When this is the case, autocorrelation can arise in two ways:

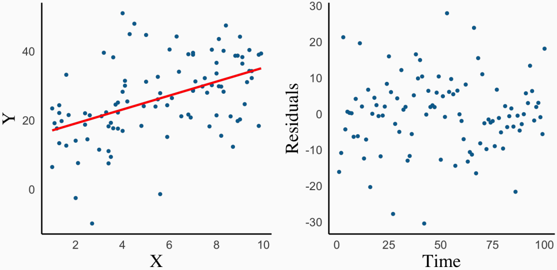
1. The autocorrelation timescale of the process we're measuring.

Temporal Autocorrelation

As biologists, we often find ourselves measuring things over time (e.g., population sizes, animal locations, CO_2 levels, reaction times, etc.).

When this is the case, autocorrelation can arise in two ways:

1. The autocorrelation timescale of the process we're measuring.

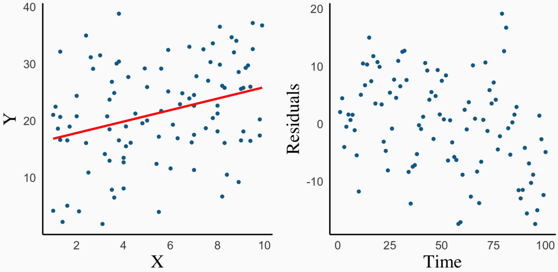


Temporal Autocorrelation

As biologists, we often find ourselves measuring things over time (e.g., population sizes, animal locations, CO_2 levels, reaction times, etc.).

When this is the case, autocorrelation can arise in two ways:

1. The autocorrelation timescale of the process we're measuring.

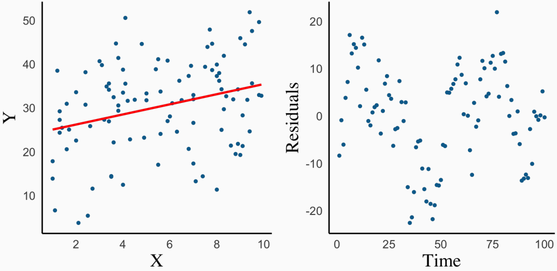


Temporal Autocorrelation

As biologists, we often find ourselves measuring things over time (e.g., population sizes, animal locations, CO_2 levels, reaction times, etc.).

When this is the case, autocorrelation can arise in two ways:

1. The autocorrelation timescale of the process we're measuring.

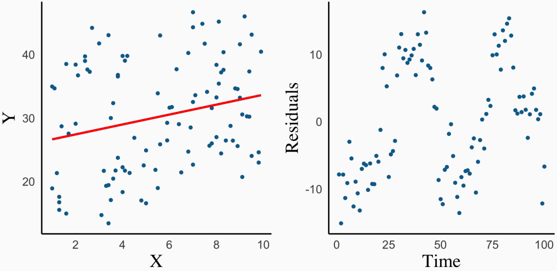


Temporal Autocorrelation

As biologists, we often find ourselves measuring things over time (e.g., population sizes, animal locations, CO_2 levels, reaction times, etc.).

When this is the case, autocorrelation can arise in two ways:

1. The autocorrelation timescale of the process we're measuring.

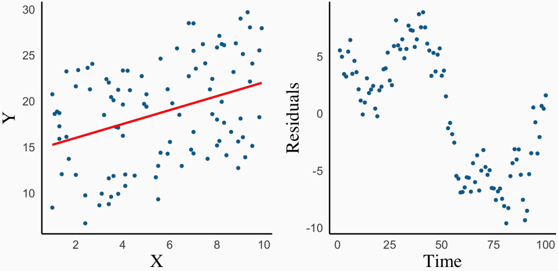


Temporal Autocorrelation

As biologists, we often find ourselves measuring things over time (e.g., population sizes, animal locations, CO_2 levels, reaction times, etc.).

When this is the case, autocorrelation can arise in two ways:

1. The autocorrelation timescale of the process we're measuring.



Temporal Autocorrelation

As biologists, we often find ourselves measuring things over time (e.g., population sizes, animal locations, CO_2 levels, reaction times, etc.).

When this is the case, autocorrelation can arise in two ways:

2. The rate/duration at which we're collecting data.

Temporal Autocorrelation

As biologists, we often find ourselves measuring things over time (e.g., population sizes, animal locations, CO_2 levels, reaction times, etc.).

When this is the case, autocorrelation can arise in two ways:

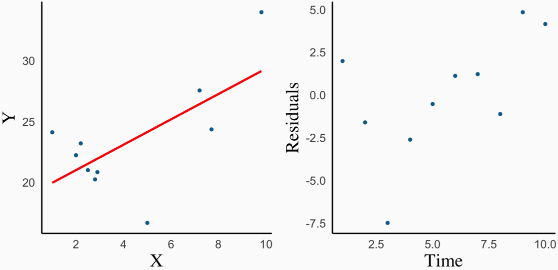
2. The rate/duration at which we're collecting data.

Temporal Autocorrelation

As biologists, we often find ourselves measuring things over time (e.g., population sizes, animal locations, CO_2 levels, reaction times, etc.).

When this is the case, autocorrelation can arise in two ways:

2. The rate/duration at which we're collecting data.

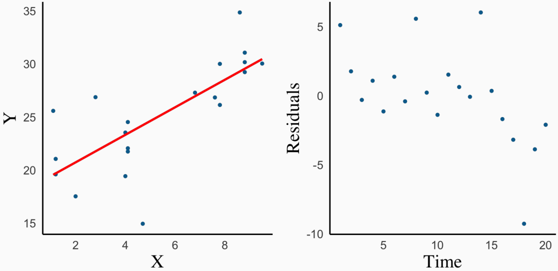


Temporal Autocorrelation

As biologists, we often find ourselves measuring things over time (e.g., population sizes, animal locations, CO_2 levels, reaction times, etc.).

When this is the case, autocorrelation can arise in two ways:

2. The rate/duration at which we're collecting data.

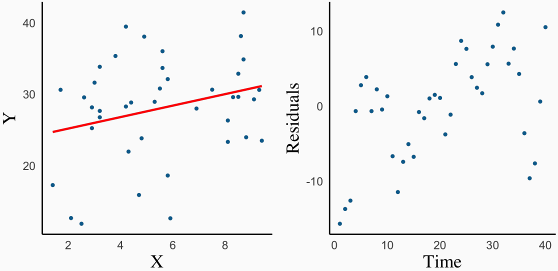


Temporal Autocorrelation

As biologists, we often find ourselves measuring things over time (e.g., population sizes, animal locations, CO_2 levels, reaction times, etc.).

When this is the case, autocorrelation can arise in two ways:

2. The rate/duration at which we're collecting data.

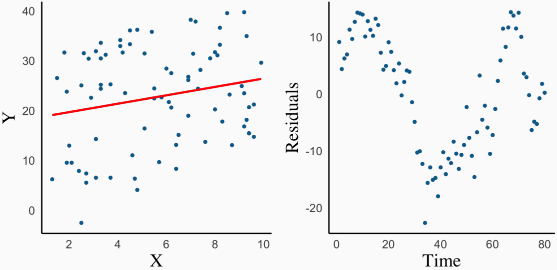


Temporal Autocorrelation

As biologists, we often find ourselves measuring things over time (e.g., population sizes, animal locations, CO_2 levels, reaction times, etc.).

When this is the case, autocorrelation can arise in two ways:

2. The rate/duration at which we're collecting data.

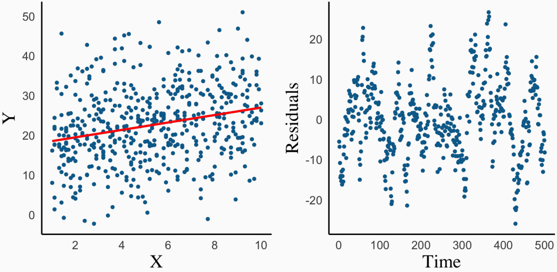


Temporal Autocorrelation

As biologists, we often find ourselves measuring things over time (e.g., population sizes, animal locations, CO_2 levels, reaction times, etc.).

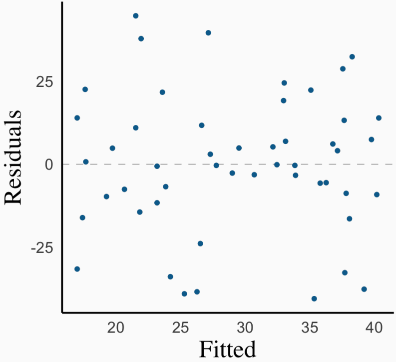
When this is the case, autocorrelation can arise in two ways:

2. The rate/duration at which we're collecting data.

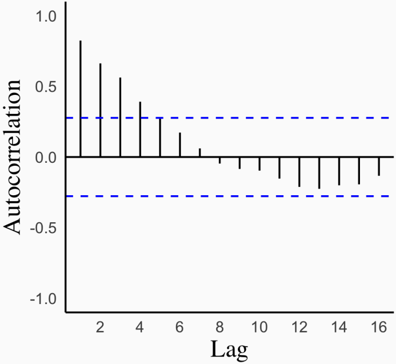


Identifying Autocorrelation

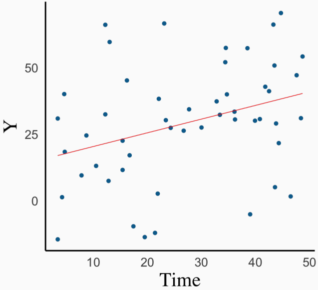
Autocorrelation can be difficult to see in a simple residuals vs. fitted plot (not designed for this purpose).



Instead we typically plot autocorrelation functions (ACFs)



Autocorrelation Function



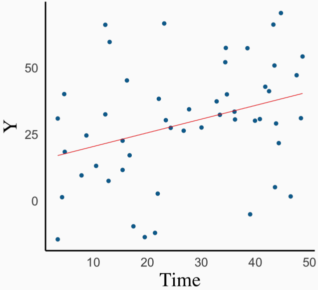
Resid_t	Resid_{t+0}
1	1
2	2
3	3
4	4
5	5
...	...
i	i

$$var_t = \frac{1}{n}\sum(\mathrm{resid_t} \times \mathrm{resid_t})$$

$$var_{t+0} = \frac{1}{n}\sum(\mathrm{resid_t} \times \mathrm{resid_{t+0}})$$

$$\mathrm{ACF}_0 = \mathrm{var}_{t+0}/\mathrm{var}_t$$

Autocorrelation Function



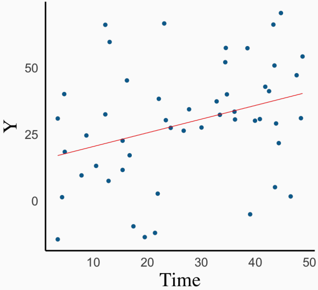
Resid_t	Resid_{t+1}
1	2
2	3
3	4
4	5
5	6
...	...
i - 1	i

$$var_t = \frac{1}{n}\sum(\mathrm{resid_t} \times \mathrm{resid_t})$$

$$var_{t+1} = \frac{1}{n}\sum(\mathrm{resid_t} \times \mathrm{resid_{t+1}})$$

$$\text{ACF}_1 = \text{var}_{t+1}/\text{var}_t$$

Autocorrelation Function



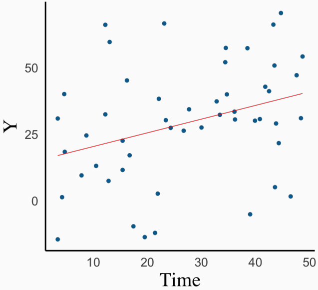
Resid_t	Resid_{t+2}
1	3
2	4
3	5
4	6
5	7
...	...
i - 2	i

$$var_t = \frac{1}{n}\sum(\mathrm{resid_t} \times \mathrm{resid_t})$$

$$var_{t+2} = \frac{1}{n}\sum(\mathrm{resid_t} \times \mathrm{resid_{t+2}})$$

$$\mathrm{ACF}_2 = \mathrm{var}_{t+2}/\mathrm{var}_t$$

Autocorrelation Function



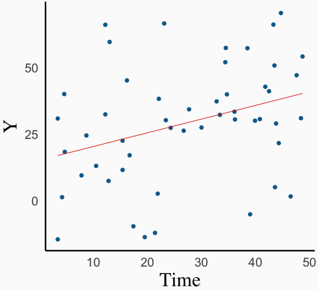
Resid$_t$	**Resid**$_{t+3}$
1	4
2	5
3	6
4	7
5	8
...	...
i - 3	i

$$var_t = \frac{1}{n}\sum(\mathrm{resid_t} \times \mathrm{resid_t})$$

$$var_{t+3} = \frac{1}{n}\sum(\mathrm{resid_t} \times \mathrm{resid_{t+3}})$$

$$\mathrm{ACF}_3 = \mathrm{var}_{t+3}/\mathrm{var}_t$$

Autocorrelation Function



Resid_t	Resid_{t+i-1}
1	i

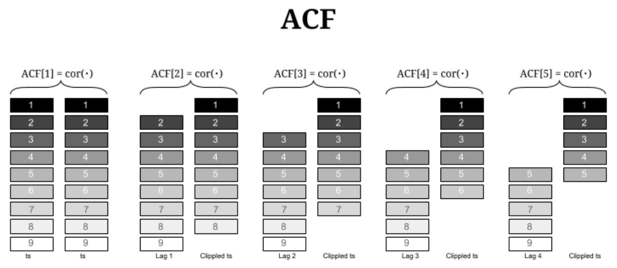
$var_t = \frac{1}{n}\sum(\mathrm{resid_t} \times \mathrm{resid_t})$

$var_{t+i-1} = \frac{1}{n}\sum(\mathrm{resid_t} \times \mathrm{resid_{t+i-1}})$

$\text{ACF}_{\text{i-1}} = \text{var}_{t+i-1}/\text{var}_t$ Continue until you've run through the whole dataset.

Autocorrelation Function cont.

Schematically, calculating the ACF looks like this:



Source: Notes on statistics and probability

What happens to the sample size as the lag increases?

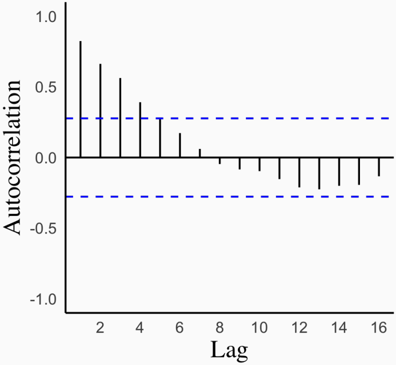
Visualising the ACF

ACF is typically used as a visual diagnostic tool.

Ranges from 1 to -1 and autocorrelation at lag $0 = 1$.

Usually accompanied by dashed lines telling you where significance lies (95% CIs).

You can do this in R via the `acf()` function.



Correcting Temporal Autocorrelation

Correcting Temporal Autocorrelation

So you find yourself with temporally autocorrelated data. What next?

Correcting Temporal Autocorrelation

‘Dealing with temporal autocorrelation’ and ‘analysing temporal trends’ are **not** the same thing.

This lecture focuses only on tools for dealing with the lack of independence associated with temporal data.

If you’re interested in analysing temporal trends you need to apply time series analysis, which we will not cover this in this course.

Temp. Autocorrelation and regression

The models we've been working with so far:

$$y_i = \beta_0 + \beta_1 \times x_i + \varepsilon_i \qquad \varepsilon_i \sim \mathcal{N}(0,\ V) \qquad V = \sigma^2 \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

The diagonal defines the variances. All 1s indicates homogeneity of variances.

The off-diagonals define the co-variances. The 0s indicate independence.

Multiplying this out would give you an $n \times 1$ matrix equal to σ^2.

$$V = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \begin{bmatrix} \sigma^2 \\ \sigma^2 \\ \vdots \\ \sigma^2 \end{bmatrix}$$

Variance-Covariance Matrix

Correcting for autocorrelation 'simply' involves identifying the autocorrelation structure of the residuals and modifying the variance-covariance matrix.

$$V = \sigma^2 \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

When the residuals are autocorrelated, the off-diagonals $\neq 0$.

$$V = \sigma^2 \underbrace{\begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}}_{\text{correlation matrix}}$$

Compound Symmetric Error Structure

The simplest case is when all the covariances are constant, non-zero value. This is referred to as ‘compound symmetry’.

The degree of correlation between residuals (ρ) is equal to $\rho = \frac{\theta}{\theta+\sigma^2}$.

$$V = \underbrace{\begin{bmatrix} \theta + \sigma^2 & \theta & \cdots & \theta \\ \theta & \theta + \sigma^2 & \cdots & \theta \\ \vdots & \vdots & \ddots & \vdots \\ \theta & \theta & \cdots & \theta + \sigma^2 \end{bmatrix}}_{\mathrm{variance-covariance\,matrix}}$$

Often too simplistic for real autocorrelation structures, but can sometimes be useful.

AR-1 Error Structure

The first order auto-regressive (AR1) structure defines a correlation structure in which the degree of correlation between two observations is proportional to the relative amount of elapsed time.

The degree of correlation between a pair of residuals is defined as $\rho^{|t-s|}$

$|t-s|$ is the absolute difference between the current time (t) and the previous time (s).

$$V = \sigma^2 \underbrace{\begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^i \\ \rho & 1 & \rho & \ddots & \vdots \\ \rho^2 & \rho & 1 & \ddots & \rho^2 \\ \vdots & \ddots & \ddots & \ddots & \rho \\ \rho^i & \cdots & \rho^2 & \rho & 1 \end{bmatrix}}_{\text{correlation matrix}}$$

E.g., If $\rho = 0.5$ and two residuals are separated by two units of time, the correlation is $0.5^2 = 0.25$.

AR-1 correlation is a useful correlation structure for ecological data.

ARMA Error Structure

The auto-regressive moving-average (ARMA) model is a general model with two parameters:

1. p: The number of autoregressive parameters.
2. q: The number of moving average parameters.

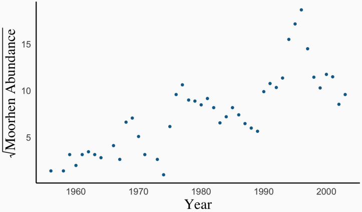
ARMA models are very flexible, but can be challenging to work with.

They can also be very slow to fit on large datasets.

Correcting Temporal Autocorrelation in R

The Hawaiian bird data

We're going to work with a dataset from Reed *et al.* (2007) to examine the abundance of moorhen (*Gallinula galeata*) on the Hawaiian Island Kauai.



The data are comprised of bird counts over years as well as annual rainfall.

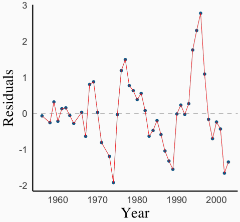
Today's starting point is the linear regression model:

$$\sqrt{\text{Birds}_i} = \beta_0 + \beta_1 \text{Rainfall}_i + \beta_2 \text{Year}_i + \varepsilon_i$$

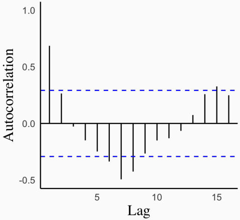
Note: The $\sqrt{}$ transformation was to clean up heteroskedasticity. We could have used the methods we learned last lecture, but we'll keep it simple today.

Autocorr. in the Hawaiian bird data

```
library(nlme)

data <- read.csv("Hawaii.csv")

data$Birds <- sqrt(data$Moorhen.Kauai)

FIT <- gls(Birds ~ Rainfall + Year, na.action = na.omit,
      data = data)

Generalized least squares fit by REML
  Model: Birds ~ Rainfall + Year
  Data: data
       AIC      BIC    logLik
  228.4798 235.4305 -110.2399

Coefficients:
                 Value Std.Error   t-value p-value
(Intercept) -477.6634  56.41907 -8.466346  0.0000
Rainfall       0.0009   0.04989  0.017245  0.9863
Year           0.2450   0.02847  8.604858  0.0000

Residual standard error: 2.608391
Degrees of freedom: 45 total; 42 residual
```



```
acf(residuals(FIT, type = "
      normalized"))
```



These data are clearly autocorrelated and the results can't be trusted.

Bird data var-cov matrix

Our model is:

$$\sqrt{Birds_i} = \beta_{\mathrm{Intercept}} + \beta_1 \times \mathrm{Rainfall}_i + \beta_2 \times \mathrm{Year}_i + \varepsilon_i \qquad \varepsilon_i \sim \mathcal{N}(0,\ V)$$

To account for the year over year autocorrelation in moorhen counts we can modify the covariances of variance-covariance matrix

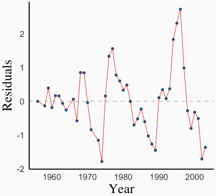
$$V = \begin{bmatrix} var_{1958} & cov_{1958,1959} & \cdots & \cdots & cov_{1958,2003} \\ cov_{1959,1958} & var_{1959} & \cdots & \ddots & \vdots \\ cov_{1960,1958} & cov_{1960,1959} & var_{1960} & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ cov_{2003,1958} & \cdots & \cdots & cov_{2003,2002} & var_{2003} \end{bmatrix}$$

Compound Symmetric Errors in R

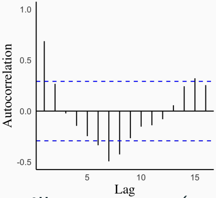
Applying autocorrelated structures to the residuals can be done using the `correlation` option in the `gls()` function.

Compound symmetric errors can be applied via the `corCompSymm()` function.

```
FIT_CompSymm <- gls(Birds ~ Rainfall + Year,
                    na.action = na.omit,
                    correlation = corCompSymm(form = ~ Year),
                    data = data)

summary(FIT_CompSymm)
...
Correlation Structure: Compound symmetry
 Formula: ~Year
 Parameter estimate(s):
         Rho
3.392348e-18
...

AIC(FIT, FIT_CompSymm)

             df      AIC
FIT           4 228.4798
FIT_CompSymm  5 230.4798
```



```
acf(residuals(FIT_CompSymm,
    type = "normalized"))
```

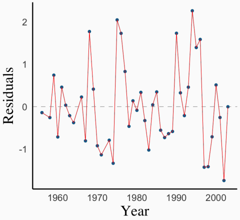


AIC shows we made the fit worse, and the residuals are still autocorr. :(

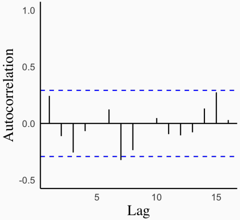
AR-1 Errors in R

AR-1 errors can be applied via the `corAR1()` function.

```
FIT_AR1 <- gls(Birds ~ Rainfall + Year,
                    na.action = na.omit,
                    correlation = corAR1(form = ~ Year),
                    data = data)

summary(FIT_AR1)
...
Correlation Structure: ARMA(1,0)
 Formula: ~Year
 Parameter estimate(s):
     Phi1
0.7734303
...

AIC(FIT, FIT_CompSymm, FIT_AR1)

             df      AIC
FIT           4 228.4798
FIT_CompSymm  5 230.4798
FIT_AR1       5 199.1394
```



```
acf(residuals(FIT_AR1, type
      = "normalized"))
```

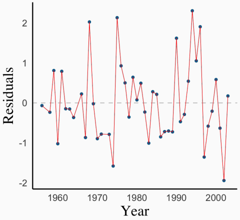


AIC shows an improvement and the residuals are no longer autocorr.

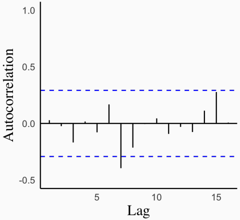
ARMA Errors in R

ARMA errors can be applied via the `corARMA()` function.

```
FIT_ARMA1 <- gls(Birds ~ Rainfall + Year,
                 na.action = na.omit,
                 correlation = corARMA(form = ~ Year,
                                       p = 2),
                 data = data)

summary(FIT_ARMA1)
...
Correlation Structure: ARMA(2,0)
 Formula: ~Year
 Parameter estimate(s):
      Phi1       Phi2
 0.9668205 -0.3220174
...

AIC(FIT, FIT_CompSymm, FIT_AR1, FIT_ARMA1)

               df      AIC
FIT             4 228.4798
FIT_CompSymm    5 230.4798
FIT_AR1         5 199.1394
FIT_ARMA1       6 196.8777
```



```
acf(residuals(FIT_ARMA1,
      type = "normalized"))
```

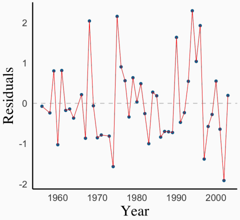


AIC shows a marginal improvement and the residuals are ok.

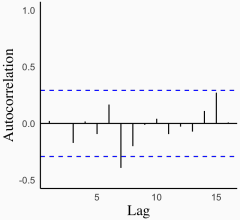
ARMA Errors in R cont.

We could also try adding a moving average term via q.

```
FIT_ARMA2 <- gls(Birds ~ Rainfall + Year,
                 na.action = na.omit,
                 correlation = corARMA(c(0.3, -0.3, 0.3),
                                       form = ~ Year,
                                       p = 2,
                                       q = 1),
                 data = data)
 summary(FIT_ARMA2)
...
Correlation Structure: ARMA(2,1)
 Formula: ~Year
 Parameter estimate(s):
       Phi1        Phi2      Theta1
 0.89422729 -0.26715887  0.08293474
...

AIC(FIT, FIT_CompSymm, FIT_AR1,
    FIT_ARMA1, FIT_ARMA2)

             df      AIC
FIT           4 228.4798
FIT_CompSymm  5 230.4798
FIT_AR1       5 199.1394
FIT_ARMA1     6 196.8777
FIT_ARMA2     7 198.8578
```



```
acf(residuals(FIT_ARMA2,
      type = "normalized"))
```



AIC is slightly worse, but the residuals are ok.

Corrected model

Original Model

```
Generalized least squares fit by REML
  Model: Birds ~ Rainfall + Year
  Data: data
       AIC      BIC    logLik
  228.4798 235.4305 -110.2399

Coefficients:
                Value Std.Error   t-value p-value
(Intercept) -477.6634  56.41907 -8.466346  0.0000
Rainfall       0.0009   0.04989  0.017245  0.9863
Year           0.2450   0.02847  8.604858  0.0000

 Correlation:
        (Intr) Ranfll
Rainfall -0.036
Year     -1.000  0.020

Residual standard error: 2.608391
Degrees of freedom: 45 total; 42 residual
```

ARMA(2,0) Model

```
Generalized least squares fit by REML
  Model: Birds ~ Rainfall + Year
  Data: data
       AIC      BIC    logLik
  196.8777 207.3037 -92.43886

Coefficients:
                Value Std.Error   t-value p-value
(Intercept) -471.8304  94.30829 -5.003064  0.0000
Rainfall      -0.0170   0.02771 -0.614301  0.5423
Year           0.2422   0.04764  5.083189  0.0000

 Correlation:
        (Intr) Ranfll
Rainfall  0.001
Year     -1.000 -0.006

Residual standard error: 2.657647
Degrees of freedom: 45 total; 42 residual

Correlation Structure: ARMA(2,0)
 Formula: ~Year
 Parameter estimate(s):
      Phi1       Phi2
 0.9668205 -0.3220174
```

Overview of Variance Structures

We covered several ways to model temporally autocorrelated data:

Type	**Covariance** ρ	**DF**	`R` **Function**
IID	0	0	`corSymm()`
Compound Symetric	$\rho = \frac{\theta}{\theta+\sigma^2}$	1	`corCompSymm()`
AR-1	$\rho^{\lvert t-s \rvert}$	1	`corAR1()`
ARMA	variable	variable	`corARMA()`

For the bird data, going from IID to AR-1 offered a big improvement, and then fine-tuning via more complicated ARMA structures resulted in only marginal improvements over AR-1. This is common in practice.

Unless there are serious issues remaining in your residuals, the pragmatic solution is to stop when you have a reasonably appropriate model.

References

Liang, M., Liu, X., Parker, I.M., Johnson, D., Zheng, Y., Luo, S., Gilbert, G.S. & Yu, S. (2019). Soil microbes drive phylogenetic diversity-productivity relationships in a subtropical forest. *Science advances*, 5, eaax5088.

Reed, J., Elphick, C., Zuur, A., Ieno, E. & Smith, G. (2007). Time series analysis of hawaiian waterbirds. In: *Analysing ecological data*. Springer, pp. 615–631.

Zuur et al. (2009) — Chapter 6