

STA 371G: Statistics and Modeling

Time Series: Autoregressive Models

Mingyuan Zhou
McCombs School of Business
The University of Texas at Austin

<http://mingyuanzhou.github.io/STA371G>

Time Series Regression... Hotel Occupancy Case

In a recent legal case, a Chicago downtown hotel claimed that it had suffered a loss of business due to what was considered an illegal action by a group of hotels that decided to leave the plaintiff out of a hotel directory.

In order to estimate the loss business, the hotel had to predict what its level of business (in terms of occupancy rate) would have been in the absence of the alleged illegal action.

In order to do this, experts testifying on behalf of the hotel use data collected before the period in question and fit a relationship between the hotels occupancy rate and overall occupancy rate in the city of Chicago. This relationship would then be used to predict occupancy rate during the period in question.

Example: Hotel Occupancy Case

SUMMARY OUTPUT

| Regression Statistics | |
|-----------------------|-----------|
| Multiple R | 0.7111011 |
| R Square | 0.5056648 |
| Adjusted R Squa | 0.48801 |
| Standard Error | 7.5055176 |
| Observations | 30 |

$$Hotel_t = \beta_0 + \beta_1 Chicago_t + \epsilon_t$$

ANOVA

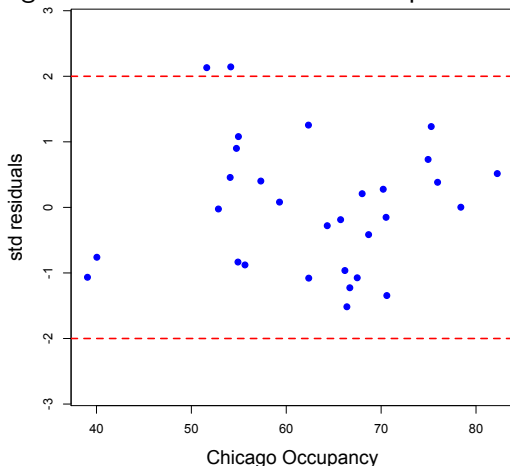
| | df | SS | MS | F | Significance F |
|------------|----|-------------|-----------|-------------|----------------|
| Regression | 1 | 1613.468442 | 1613.4684 | 28.64172598 | 1.06082E-05 |
| Residual | 28 | 1577.318225 | 56.332794 | | |
| Total | 29 | 3190.786667 | | | |

| | Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% |
|------------|--------------|----------------|-----------|-------------|--------------|-------------|
| Intercept | 16.135666 | 8.518889357 | 1.8941044 | 0.068584205 | -1.314487337 | 33.5858198 |
| ChicagoInd | 0.7161318 | 0.133811486 | 5.3517965 | 1.06082E-05 | 0.442031445 | 0.990232246 |

- In the month after the omission from the directory the Chicago occupancy rate was 66. The plaintiff claims that its occupancy rate should have been $16 + 0.71 \cdot 66 = 62$.
- It was actually 55!! The difference added up to a big loss!!
- Under this model, what's the probability for $Hotel_{31} \leq 55$?

Example: Hotel Occupancy Case

A statistician was hired by the directory to access the regression methodology used to justify the claim. As we should know by now, the first thing he looked at was the residual plot...



Looks fine. However...

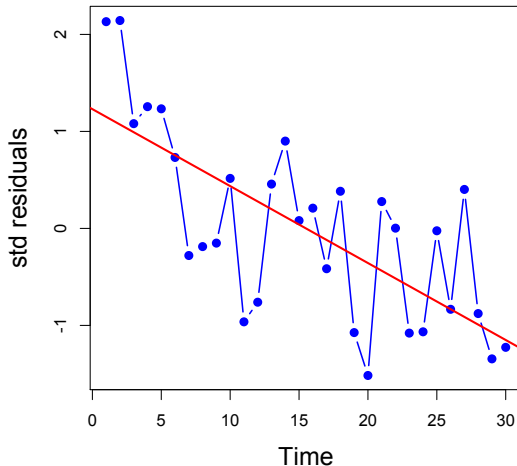
Example: Hotel Occupancy Case

... this is a *time series regression*, as we are regressing one time series on another.

In this case, we should also check whether or not the residuals show some temporal pattern.

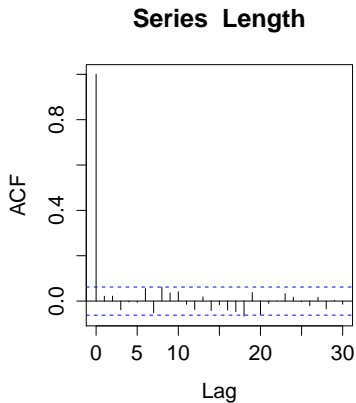
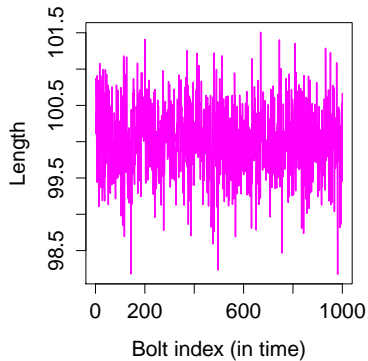
If our model is correct the residuals should look iid normal over time.

Example: Hotel Occupancy Case

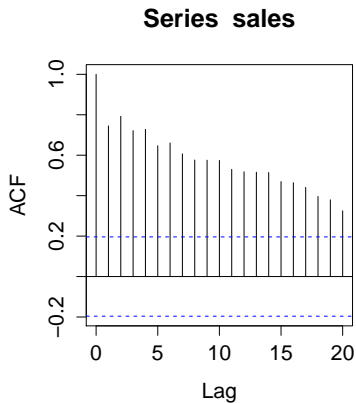
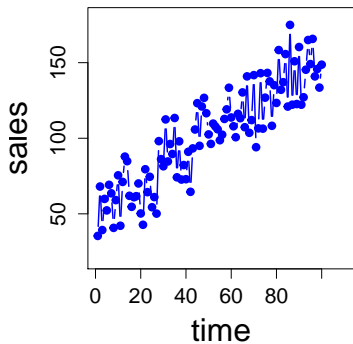


Does this look iid to you? Can you guess what the red line represent?

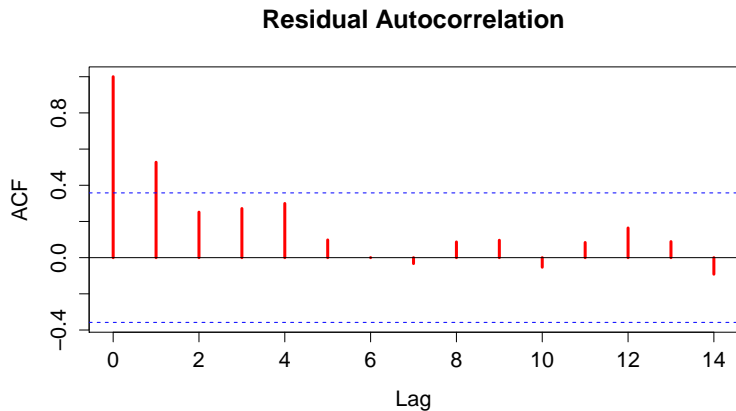
Autocorrelation



Autocorrelation



Example: Hotel Occupancy Case



Example: Hotel Occupancy Case

It looks like part of hotel occupancy (y) not explained by the Chicago downtown occupancy (x) is moving down over time. We can try to control for that by adding a trend factor to our model...

$$Hotel_t = \beta_0 + \beta_1 Chicago_t + \beta_2 t + \epsilon_t$$

SUMMARY OUTPUT

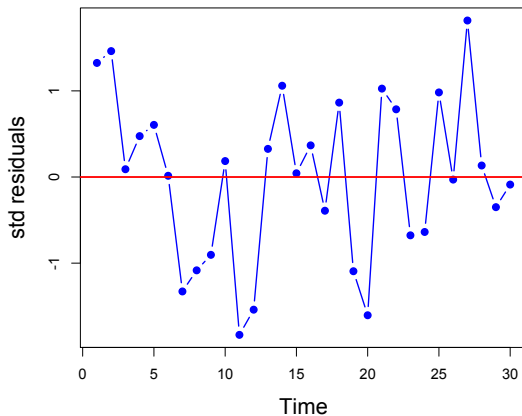
| <i>Regression Statistics</i> | |
|------------------------------|-------------|
| Multiple R | 0.869389917 |
| R Square | 0.755838827 |
| Adjusted R Squ | 0.737752815 |
| Standard Error | 5.37162026 |
| Observations | 30 |

ANOVA

| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>Significance F</i> |
|------------|-----------|-------------|-----------|-------------|-----------------------|
| Regression | 2 | 2411.720453 | 1205.86 | 41.79134652 | 5.41544E-09 |
| Residual | 27 | 779.0662139 | 28.8543 | | |
| Total | 29 | 3190.786667 | | | |

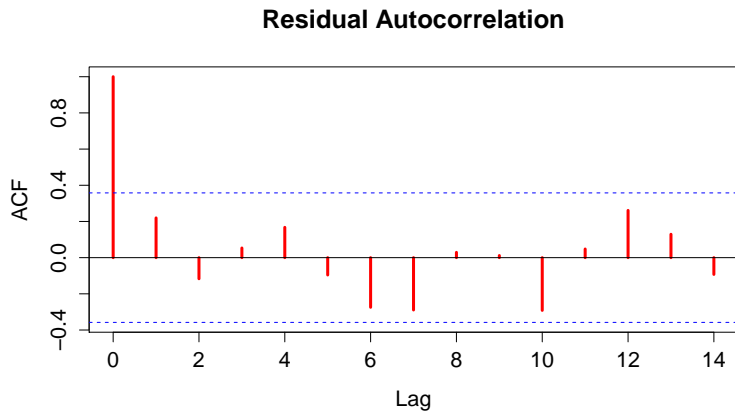
| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> | <i>Lower 95%</i> | <i>Upper 95%</i> |
|------------|---------------------|-----------------------|---------------|----------------|------------------|------------------|
| Intercept | 26.69391108 | 6.418837165 | 4.158683 | 0.000290493 | 13.52354525 | 39.8642769 |
| ChicagoInd | 0.69523791 | 0.095849831 | 7.253408 | 8.41391E-08 | 0.498570304 | 0.89190552 |
| t | -0.596476666 | 0.113404099 | -5.259745 | 1.51653E-05 | -0.82916265 | -0.3637907 |

Example: Hotel Occupancy Case



Much better!! What is the slope of the red line?

Example: Hotel Occupancy Case



Example: Hotel Occupancy Case

Okay, what happened?!

Well, once we account for the downward trend in the occupancy of the plaintiff, the prediction for the occupancy rate is

$$26 + 0.69 * 66 - 0.59 * 31 = 53.25$$

What do we conclude?

Example: Hotel Occupancy Case

Take away lessons...

- ▶ When regressing a time series on another, always check the residuals as a time series
- ▶ What does that mean... plot the residuals over time. If all is well, you should see no patterns, i.e., they should behave like iid normal samples.

Example: Hotel Occupancy Case

Question

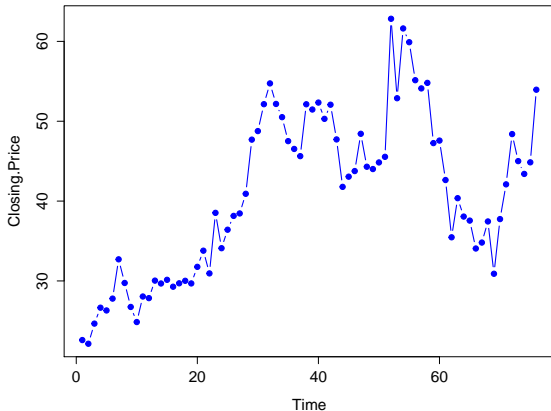
- ▶ What if we were interested in predicting the hotel occupancy ten years from now?? We should compute

$$26 + 0.69 * 66 - 0.59 * 150 = -16.96$$

- ▶ Would you trust this prediction? Could you defend it in court?
- ▶ Remember: always be careful with extrapolating relationships!

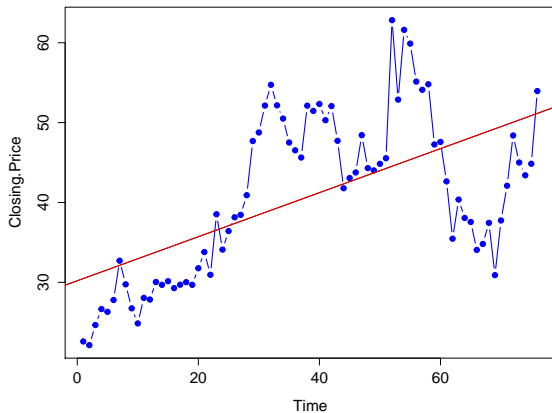
Example: Monthly Stock Closing Prices

How to model this time series?



Example: Monthly Stock Closing Prices

Let's first model the trend using $Y_t = \beta_0 + \beta_1 t + \epsilon_t$



Example: Monthly Stock Closing Prices

The following summary looks fine

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) | |
|-------------|----------|------------|---------|----------|-----|
| (Intercept) | 30.19762 | 1.90827 | 15.825 | < 2e-16 | *** |
| Time | 0.27577 | 0.04307 | 6.404 | 1.24e-08 | *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

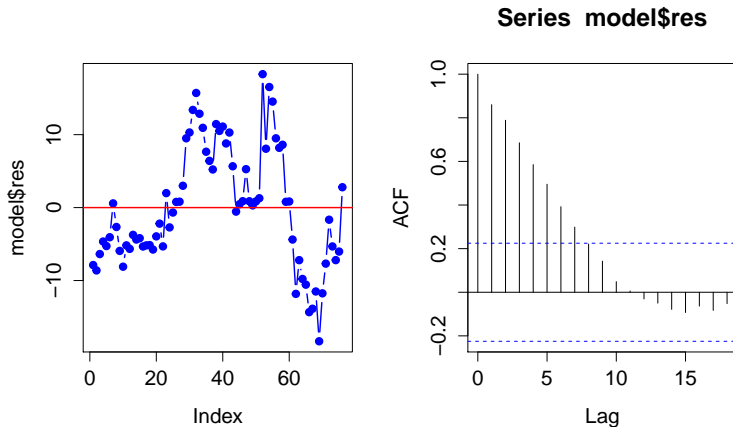
Residual standard error: 8.236 on 74 degrees of freedom

Multiple R-squared: 0.3565, Adjusted R-squared: 0.3479

F-statistic: 41 on 1 and 74 DF, p-value: 1.245e-08

Example: Monthly Stock Closing Prices

But let's look at the residual time series and its autocorrelation



Random Walk Model

$$Y_t = Y_{t-1} + \mu + \epsilon_t$$

- ▶ The time series itself is not random
- ▶ The differences between consecutive times are random

$$Y_t - Y_{t-1} = \mu + \epsilon_t$$

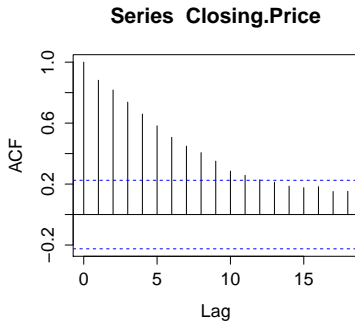
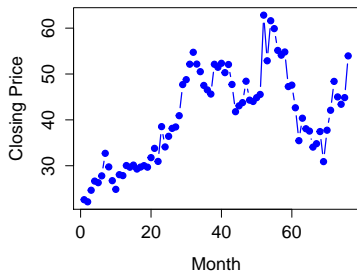
Random Walk Model

$$Y_t = Y_{t-1} + \mu + \epsilon_t$$

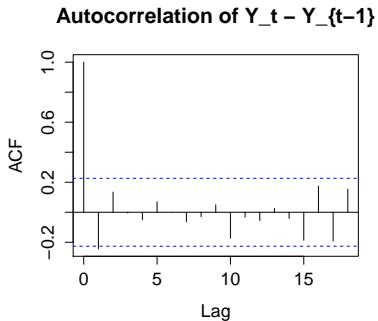
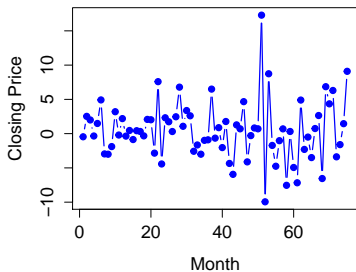
Analyzing monthly stock closing prices



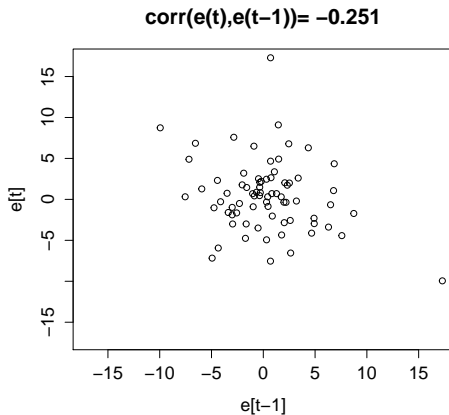
Random Walk Model



Random Walk Model



Random Walk Model



Can we do better?

The AR(1) Model

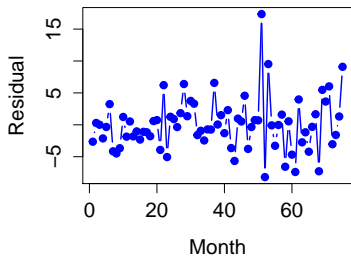
A simple way to model dependence over time in with the autoregressive model of order 1...

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \epsilon_t$$

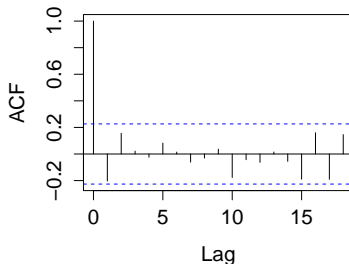
- ▶ What is the mean of Y_t for a given value of Y_{t-1} ?
- ▶ If the model successfully captures the dependence structure in the data then the residuals should look iid.
- ▶ Remember: if our data is collected in time, we should always check for dependence in the residuals...

Example: Monthly Stock Closing Prices

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \epsilon_t$$



Autocorrelation of Residuals



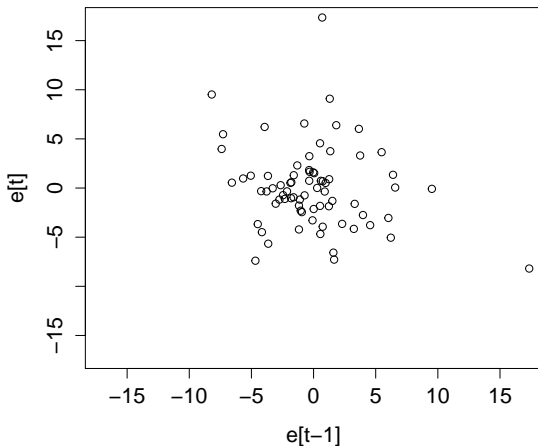
Example: Monthly Stock Closing Prices

| Regression Statistics | | | | | | |
|---|---------------------|-----------------------|-------------|------------|----------------|----------------|
| <i>R</i> | 0.91171 | | | | | |
| <i>R Square</i> | 0.83121 | | | | | |
| <i>Adjusted R Square</i> | 0.8289 | | | | | |
| <i>Standard Error</i> | 4.15447 | | | | | |
| <i>Total Number Of Cases</i> | 75 | | | | | |
| Closing Price [t] = 4.4009 + 0.9020 * Closing Price [t-1] | | | | | | |
| ANOVA | | | | | | |
| | <i>d.f.</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>p-level</i> | |
| <i>Regression</i> | 1. | 6,204.67064 | 6,204.67064 | 359.49056 | 0.E+0 | |
| <i>Residual</i> | 73. | 1,259.95229 | 17.25962 | | | |
| <i>Total</i> | 74. | 7,464.62293 | | | | |
| | <i>Coefficients</i> | <i>Standard Error</i> | <i>LCL</i> | <i>UCL</i> | <i>t Stat</i> | <i>p-level</i> |
| Intercept | 4.40093 | 1.99197 | 0.43093 | 8.37092 | 2.20933 | 0.03029 |
| Closing Price [t-1] | 0.90199 | 0.04757 | 0.80718 | 0.99681 | 18.96024 | 0.E+0 |

Example: Monthly Stock Closing Prices

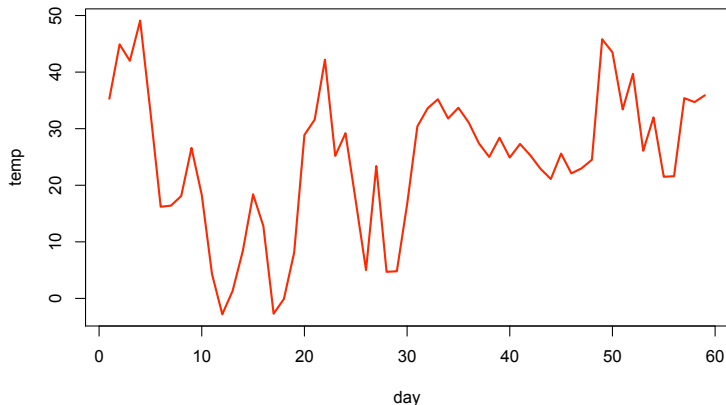
$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \epsilon_t$$

$$\text{corr}(\epsilon_t, \epsilon_{t-1}) = -0.211$$



Examples: Temperatures

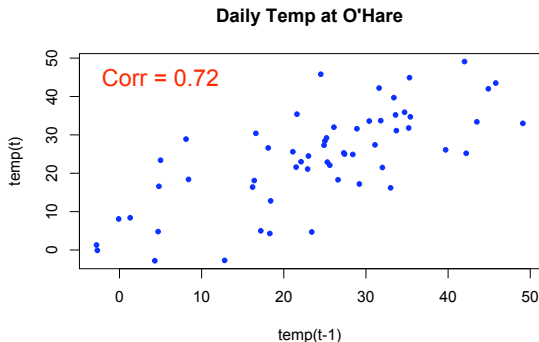
Now you need to predict tomorrow's temperature at O'Hare from (Jan-Feb).



Does this look iid? If it is iid, tomorrow's temperatures should not depend on today's... does that make sense?

Checking for Dependence

To see if Y_{t-1} would be useful for predicting Y_t , just plot them together and see if there is a relationship.



Correlation between Y_t and Y_{t-1} is called **autocorrelation**.

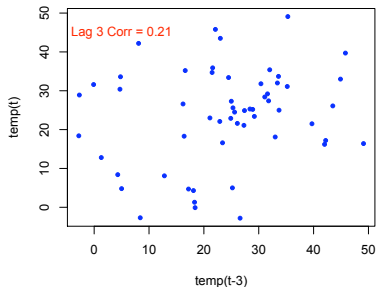
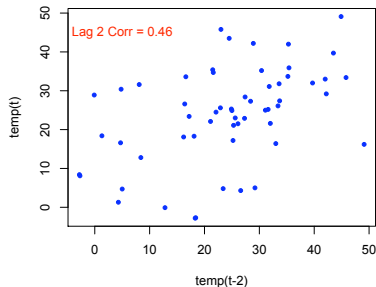
Checking for Dependence

You need to create a “lagged” variable $temp_{t-1}$... the data looks like this:

| t | temp(t) | temp(t-1) |
|-----|---------|-----------|
| 1 | 42 | 35 |
| 2 | 41 | 42 |
| 3 | 50 | 41 |
| 4 | 19 | 50 |
| 5 | 19 | 19 |
| 6 | 20 | 19 |
| ... | ... | ... |

Checking for Dependence

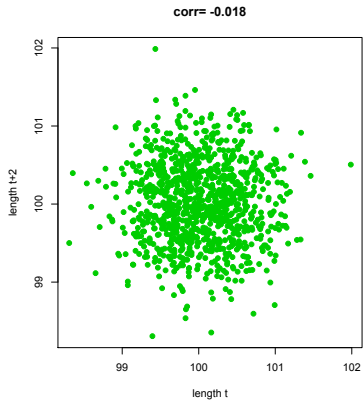
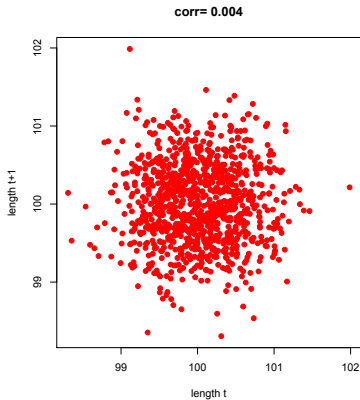
We can plot Y_t against Y_{t-L} to see **L-period lagged relationships**.



- ▶ It appears that the correlation is getting weaker with increasing L .
- ▶ **How can we test for this dependence?**

Checking for Dependence

Back to the “length of a bolt” example. When things are not related in time we should see...



The AR(1) Model

Again, the regression tool is our friend here... (Why?)

SUMMARY OUTPUT

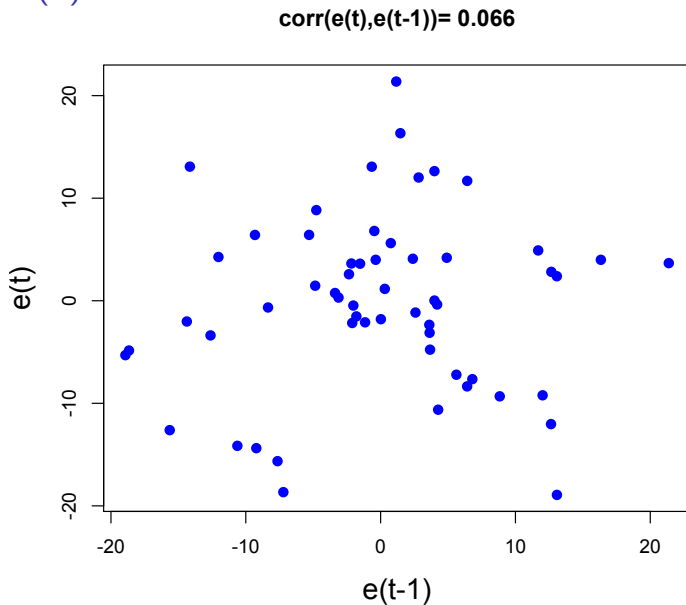
| <i>Regression Statistics</i> | |
|------------------------------|-------------|
| Multiple R | 0.722742583 |
| R Square | 0.522356842 |
| Adjusted R Sq | 0.5138275 |
| Standard Error | 8.789861051 |
| Observations | 58 |

ANOVA

| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>Significance F</i> |
|------------|-----------|-------------|-------------|-------------|-----------------------|
| Regression | 1 | 4731.684433 | 4731.684433 | 61.24233673 | 1.49699E-10 |
| Residual | 56 | 4326.652809 | 77.2616573 | | |
| Total | 57 | 9058.337241 | | | |

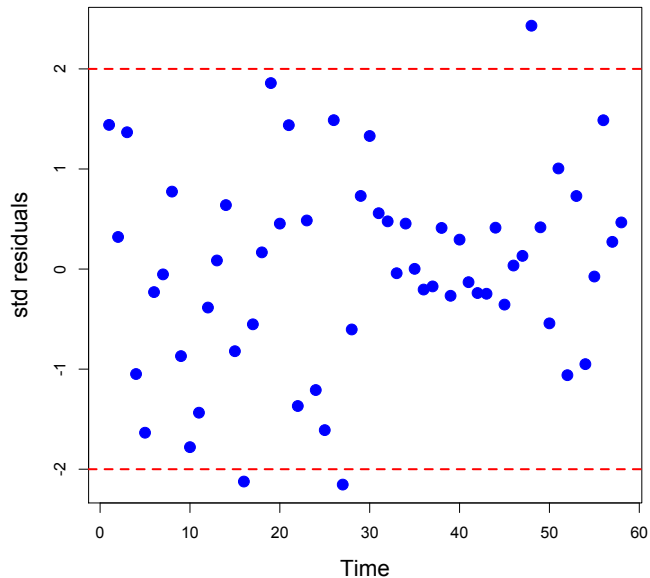
| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> | <i>Lower 95%</i> | <i>Upper 95%</i> |
|--------------|---------------------|-----------------------|---------------|----------------|------------------|------------------|
| Intercept | 6.705800085 | 2.516614758 | 2.664611285 | 0.010050177 | 1.664414964 | 11.74718521 |
| X Variable 1 | 0.723288866 | 0.092424243 | 7.825748317 | 1.49699E-10 | 0.53814086 | 0.908436873 |

The AR(1) Model



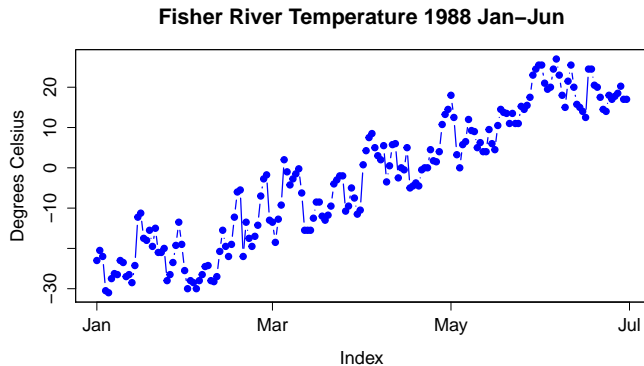
No dependence left!

The AR(1) Model



Again, looks good...

Example: Fisher River Temperature 1988 January to June



Example: Fisher River Temperature 1988 January to June

$$Y_t = \beta_0 + \beta_1 t + \epsilon_t$$

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|------------|------------|---------|------------|
| (Intercept) | -29.105929 | 0.860407 | -33.83 | <2e-16 *** |
| Time | 0.288914 | 0.008155 | 35.43 | <2e-16 *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

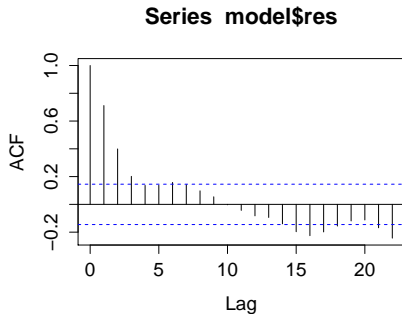
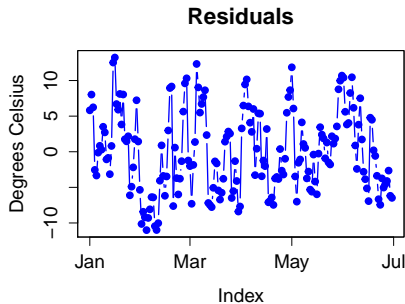
Residual standard error: 5.78 on 180 degrees of freedom

Multiple R-squared: 0.8746, Adjusted R-squared: 0.8739

F-statistic: 1255 on 1 and 180 DF, p-value: < 2.2e-16

Example: Fisher River Temperature 1988 January to June

$$Y_t = \beta_0 + \beta_1 t + \epsilon_t$$



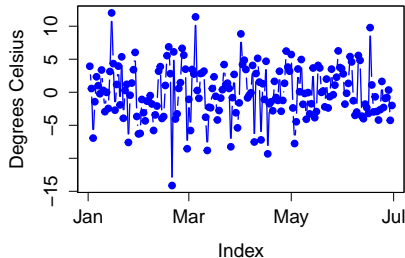
Example: Fisher River Temperature 1988 January to June

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 t + \epsilon_t$$

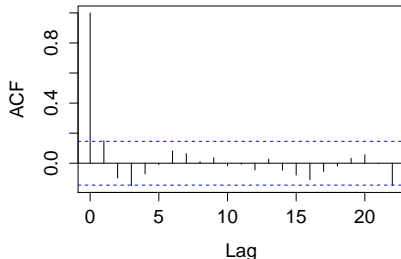
Example: Fisher River Temperature 1988 January to June

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 t + \epsilon_t$$

Residuals



Series model\$res



Example: Fisher River Temperature 1988 January to June

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 t + \epsilon_t$$

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) | |
|---------------|----------|------------|---------|----------|-----|
| (Intercept) | -8.03628 | 1.64254 | -4.893 | 2.22e-06 | *** |
| Fisher[1:181] | 0.71663 | 0.05235 | 13.689 | < 2e-16 | *** |
| Time | 0.08209 | 0.01624 | 5.054 | 1.07e-06 | *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

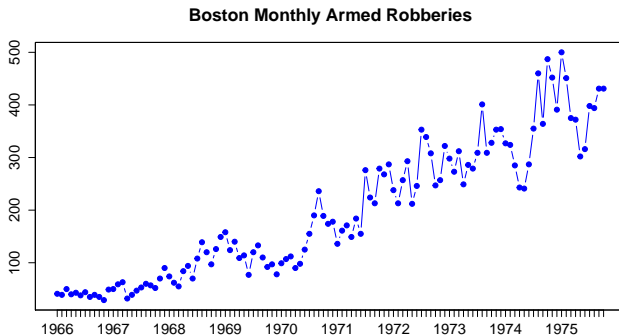
Residual standard error: 4.045 on 178 degrees of freedom

Multiple R-squared: 0.9387, Adjusted R-squared: 0.938

F-statistic: 1363 on 2 and 178 DF, p-value: < 2.2e-16

Example: Monthly Boston Armed Robberies

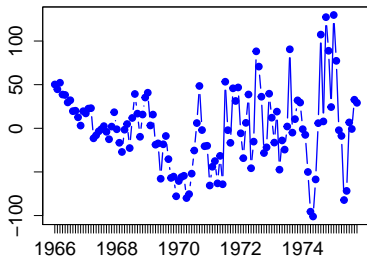
Jan.1966-Oct.1975



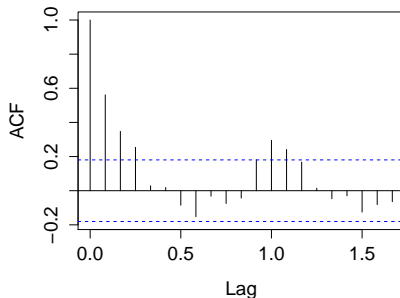
Example: Boston Armed Robberies

$$Y_t = \beta_0 + \beta_1 t + \epsilon_t$$

Residuals

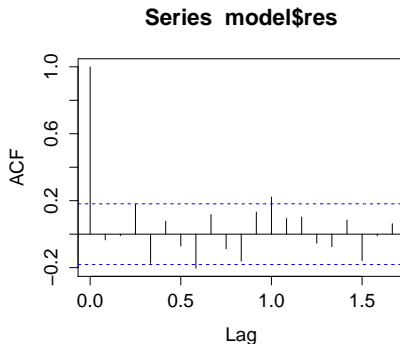
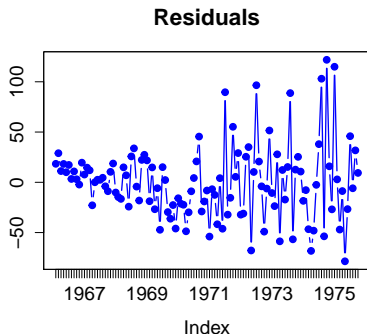


Series model\$res



Example: Boston Armed Robberies

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 t + \epsilon_t$$



Are the model assumptions violated?

Can we do better?

Example: Boston Armed Robberies

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 t + \epsilon_t$$

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|--|----------|------------|---------|--------------|
| (Intercept) | -4.06758 | 6.89469 | -0.590 | 0.556 |
| BostonRobbery[1:length(BostonRobbery) - 1] | 0.56397 | 0.07691 | 7.333 | 3.52e-11 *** |
| Time | 1.56124 | 0.28733 | 5.434 | 3.17e-07 *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 36.7 on 114 degrees of freedom

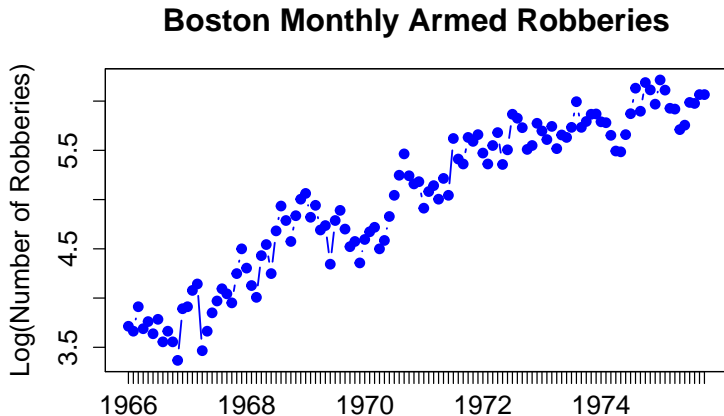
Multiple R-squared: 0.9189, Adjusted R-squared: 0.9175

F-statistic: 645.9 on 2 and 114 DF, p-value: < 2.2e-16

Interpretation?

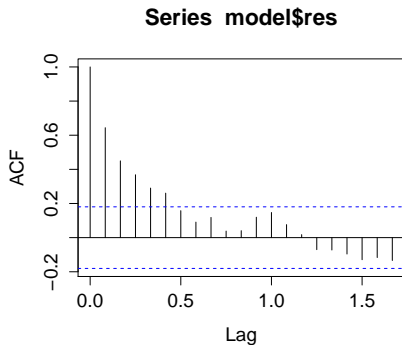
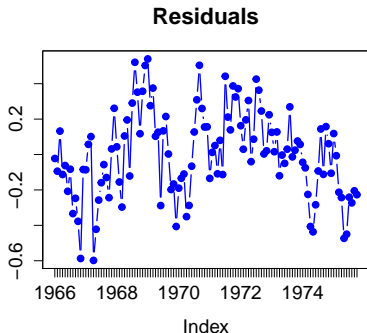
Example: Boston Armed Robberies

Let's take the log transformation of the number of armed robberies



Example: Boston Armed Robberies

$$\text{Log}(Y_t) = \beta_0 + \beta_1 t + \epsilon_t$$



Can we do better?

Example: Boston Armed Robberies

$$\text{Log}(Y_t) = \beta_0 + \beta_1 t + \epsilon_t$$

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) | |
|-------------|-----------|------------|---------|----------|-----|
| (Intercept) | 3.7142003 | 0.0457093 | 81.26 | <2e-16 | *** |
| Time | 0.0218550 | 0.0006667 | 32.78 | <2e-16 | *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2467 on 116 degrees of freedom

Multiple R-squared: 0.9026, Adjusted R-squared: 0.9017

F-statistic: 1075 on 1 and 116 DF, p-value: < 2.2e-16

Interpretation?

Example: Boston Armed Robberies

$$\text{Log}(Y_t) = \beta_0 + \beta_1 t + \epsilon_t$$

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) | |
|-------------|-----------|------------|---------|----------|-----|
| (Intercept) | 3.7142003 | 0.0457093 | 81.26 | <2e-16 | *** |
| Time | 0.0218550 | 0.0006667 | 32.78 | <2e-16 | *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2467 on 116 degrees of freedom

Multiple R-squared: 0.9026, Adjusted R-squared: 0.9017

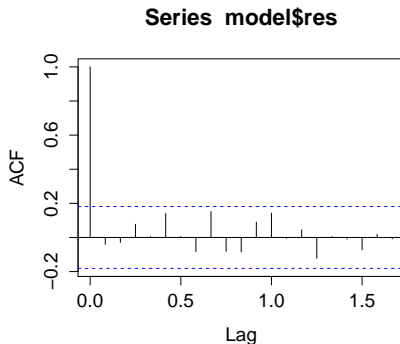
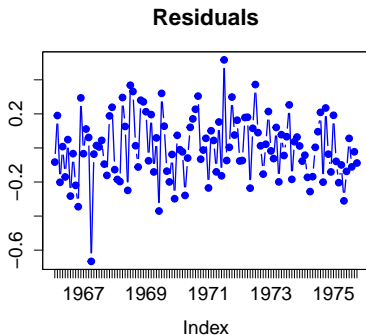
F-statistic: 1075 on 1 and 116 DF, p-value: < 2.2e-16

Interpretation?

The number of Armed Robberies increased by about 2.2% every month!

Example: Boston Armed Robberies

$$\text{Log}(Y_t) = \beta_0 + \beta_1 \text{Log}(Y_{t-1}) + \beta_2 t + \epsilon_t$$



Much better!

Example: Boston Armed Robberies

$$\text{Log}(Y_t) = \beta_0 + \beta_1 \text{Log}(Y_{t-1}) + \beta_2 t + \epsilon_t$$

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) | |
|---|----------|------------|---------|----------|-----|
| (Intercept) | 1.329217 | 0.268444 | 4.952 | 2.57e-06 | *** |
| log(BostonRobbery[1:length(BostonRobbery) - 1]) | 0.648915 | 0.071723 | 9.047 | 4.57e-15 | *** |
| Time | 0.007598 | 0.001658 | 4.582 | 1.19e-05 | *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1898 on 114 degrees of freedom

Multiple R-squared: 0.9419, Adjusted R-squared: 0.9409

F-statistic: 924.5 on 2 and 114 DF, p-value: < 2.2e-16