WEEK 3 – EXPLANATORY ANALYSIS AND FORECASTING

ASSIGNMENT 1 IS DUE

Fisher Exact Test

For Loops

- For loops let you repeat code for each element of a vector
- Suppose you go to subway and order a meatball sub with lettuce, tomato, green pepper, and pickles
- The sandwich artist will put on each of these ingredients in sequentially
- As a for loop, it would could look like this:

```
#For loop example
finalSub = c('','','','')
ingredientList = c('tomato','spinach','green
pepper','pickles')

for(ingredientNum in c(1:4)){
     finalSub[ingredientNum] =
ingredientList[ingredientNum]
}
```

For Loop Demonstration – Line-by-Line

- `for(ingredientNum in c(1:4)){`
  - The for loop goes through each value in the vector c(1:4)
  - It assigns that value to the variable 'ingredientNum'

- `    finalSub[ingredientNum] = ingredientList[ingredientNum]`
  - Assigns the value in element `ingredientNum` of vector `ingredientList` to element `ingredientNum` of vector `finalSub`

- }
  - Close the for loop

For Loop - Demonstration

- To print out the numbers from 1 to 1000
  - `for(i in 1:1000){`
  - `   print(i)`
  - `}`
- Populate a vector called dpiGreaterThan1000
  - `#For loop example 3`
  - `LifeCycleSavings$dpiGreaterThan1000 = FALSE`
  - `for(i in 1:50){`
  - `   LifeCycleSavings$dpiGreaterThan1000[i] = LifeCycleSavings$dpi[i] > 1000`
  - `}`

Loops - Readability

- Loops can be nested multiple times
- To keep code clear, it is very important to use consistent indentation

- ```
for(i in 1:5){
        print(i)
        for(j in 1:5){
                print(i*j)
        }
}
```
```

Loop Efficiency

- Given the scale of the data, we have to be aware of efficiency concerns
- Suppose we have a database where each row corresponds to a purchase made by a consumer
- Compare the following two bits of pseudocode to calculate whether that consumers owns a car

Loop Efficiency

- #For each row in the database
  - #Find the consumer in this row
  - #Find corresponding entry in car database
  - #Assign isCarOwner variable in this row

- #For each consumer in the database
  - #Find all rows that are the consumers
  - #Find corresponding entry in car database
  - #Assign isCarOwner variable for all rows

Loops - Review

- Loops perform an operation many times – once for each item in a vector
  - The vector can be made of any datatype

- Loops have a specific syntax that needs to be followed
- Loops should be indented properly to be readable

Loops - Quiz

- What do the following code segments do?
- cycleVec = c(9, 7, 5, 3)

- for(i in cycleVec){
```
for(i in cycleVec){
        print(i)
}
```

- for(j in 1:5) {
```
for(j in 1:5) {
        for(i in cycleVec){
                print(i)
        }
}
```

Pseudocode

- Key to writing pseudocode is very carefully imaging each step of a process
- If you’re having trouble coding something, break it down into smaller steps

- What might be the pseudocode for a function that takes the mean of a numeric vector?

Pseudo-code and for loops

- Think of a for loop as a new employee
- If you show it how to do a task once, it can then do it many times
- If you can't figure out how to to each thing, figure out how to do the first thing
- Simple Example: how might we print the numbers from 1 to 1000?
  - Start with printing '1'
  - Then use the loop, replacing '1' with the looping variable

Pseudo-code and for loops

- `i = 1`
- `print(1)`

- ```
  for(i in 1:1000){
      print(i)
  }
  ```
  ```

Pseudo-code and for loops

- Suppose you have a database full of reviews for various products
- How do you find the most recommended review for each product?
- Instead, ‘How do you find the most recommended review for the first product?’

Pseudo-code and for loops

- Lets try another for loop – this time you write the pseudo-code too
- How do we create a database with the first review for each product?
- First work on how to do it for the first product.

Combining Loops and Regressions

- What if we want to see how units purchased reacts to prices at each store?
  - Some stores might have more price sensitive consumers

- Quiz: write pseudocode for this
  - Hint: try to write code for one store, then use a for loop to generalize to multiple stores

- #Create blank data to store
- priceCoefByStore = rep(NA,max(consumerData$STOREIdentifier))

- #Start a for loop – for each store number
- for(storeNum in 1:max(consumerData$STOREIdentifier)){
- #Take the right subset of the data
- currentStoreData = subset(consumerData,consumerData$STOREIdentifier==storeNum)

- #Run the regression on this subset
- currentStoreLM = lm(units~price,data=currentStoreData)
- #Get the coefficient from the store data
- priceCoefByStore[storeNum] = currentStoreLM$coefficients[2]
- }

Plots in R - Loops

- Quiz: Try and write pseudocode to plot the units versus price plot for each store in the data
  - Hint similar pattern as before – figure out how to do it for one store, then generalize

Plots in R - Loops

- `for(storeNum in 1:max(consumerData$STOREIdentifier)){`
- `      storeData = subset(consumerData,consumerData$STOREIdentifier==storeNum)`
- `      pdf(paste('plotStore',storeNum,'.pdf‘,sep=‘’))`
- `      plot(storeData$units~storeData$price)`
- `      dev.off()`
- `}`

Packages - Basics

- Packages are add-ons to R
- You can install them from the R console, with R code
- Thousands of packages!
- To install and load a package, use `install.packages` and `library` functions
- For example, to install and load the 'tm' package, write
  - `install.packages('tm')`
  - `library('tm')`

Packages – Sample Code

- Using these functions is preferred to loading the data using R-Studio
- Programs work best with as little manual steps as possible
- If you manually load packages and datasets with R-Studio, you have to do that again every time you run the script
- If you use code, you don't have to do anything

Packages – World Cloud of Cat Toy Reviews

- `fullReviewDB = read.csv('catToyReviews.csv')`

- `install.packages('tm')`
- `install.packages('SnowballC')`
- `install.packages('wordcloud')`

- `library(tm)`
- `library(SnowballC)`
- `library(wordcloud)`

- `reviewCorpus = Corpus(VectorSource(fullReviewDB$review.text))`
- `reviewCorpus = tm_map(reviewCorpus, removeWords, stopwords('english'))`
- `wordcloud(reviewCorpus, max.words = 100, random.order = FALSE)`

Packages – Help Files

- When a package is loaded, it loads help files for all associated functions
- Check the help file for `wordcloud` by typing `?wordcloud`

Causal Analysis Practice – Beer: Cans and Bottles

Causal Analysis Practice – Beer: Cans and Bottles

- Beer Brewery is considering whether to sell their beer in Bottles
- They currently sell in cans

1. What is the problem with the following regression: Sales ~ (isCan) + (isBottle) + (isCan*isBottle)
2. What experiment could help determine the additional sales
3. How might you approximate an experiment using real world data?

Main Categories of Data Analysis

- Explanatory: *Summarize the data*
  - Exploratory
  - Descriptive

- Predictive: *Predict the data*
  - Statistical
  - “Futurology”

- Causal: *Change the data*
  - Econometric

Main Categories of Data Analysis – What correlation means means

- Explanatory: *Summarize the data*
  - An interesting relationship
- Predictive: *Predict the data*
  - A potentially good predictor
- Causal: *Change the data*
  - Maybe nothing?

What is a metric?

- A metric is a numeric value that can be used to evaluate something
- In our case, metrics are used to evaluate models
- Different kinds of analysis have different metrics

Main Categories of Data Analysis: Metrics

- Explanatory: *Summarize the data*
  - R-Squared
  - Adjusted R-Squared
  - AIC
  - BIC

- Predictive: *Predict the data*
  - Predictive Validity

- Causal: *Change the data*
  - Data Source

Causal Analysis Practice: Violent Movies

Causal Analysis Practice: Violent Movies

- Research Question: Do violent movies lead to an increase in crime?
- What’s wrong with an experimental approach here?
- What’s wrong with a survey approach here?

Causal Analysis Practice: Violent Movies

- What variables do you think would need to be controlled for when analyzing this question using real world data?

EXPLANATORY ANALYSIS

Explanatory Analysis

- Used to
  - Find interesting trends in the data
  - Summarize the data
  - Make the data more perceptible
  - Basically, interested in correlations
- You already use explanatory analysis to evaluate data
  - Scatterplots
  - Correlations
  - Matrices of both
- To use these to evaluate models, we have to find the right metrics

R-squared

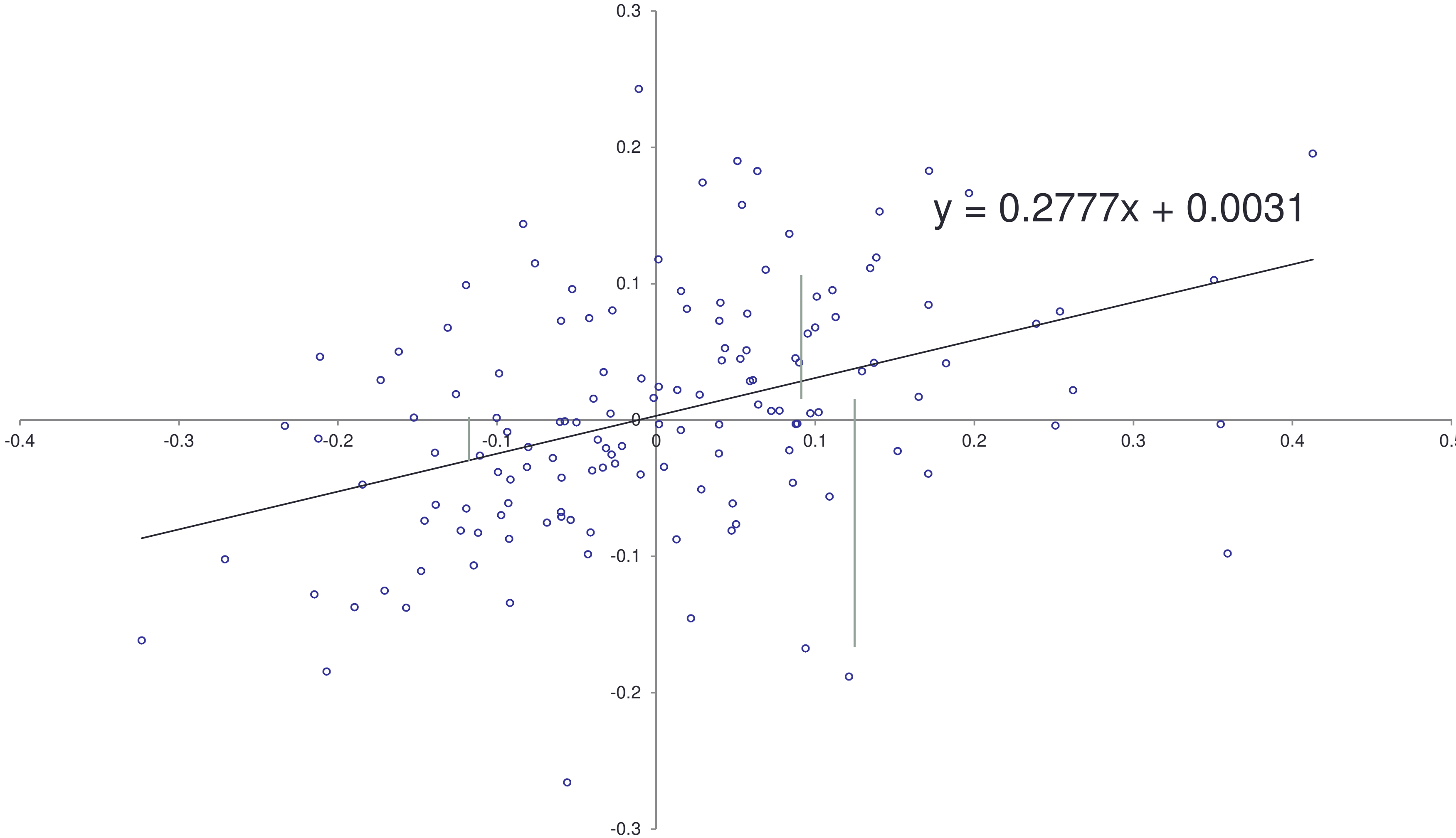
- The goodness-of-fit measure, R^2, is a measure of the extent to which the variation of the dependent variable is explained by the explanatory variable(s).

- $0 < R^2 < 1$.

- R^2 close to 1 indicate good explanatory power.

- $R^2 = 1 - \frac{sum\ of\ squared\ errors}{sum\ of\ deviations\ from\ mean}$

- It measures the amount of variation in the data that can be explained by the regression.



y = 0.2777x + 0.0031
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5

Why can R-Squared be Bad?

- How can we get 100% R-Squared?
  - ```
    y = c(1,5,3,10)
    x = c(1,2,3, 4)
    basicLM = lm(y~x)
    factorLM = lm(y~factor(x))
    ```

- Check R-squared on both
    ```

Why can R-Squared be Bad?

- What happens if we add random data
  - ```
    y = c(1,5,3,10)
    x = c(1,2,3, 4)
    #rand = ????
    baseLM = lm(y~x)
    factorLM = lm(y~x+rand)
    ```

- Check R-squared on both – which is bigger?
    ```

Why can R-Squared be Bad?

- Fundamental problem is that R-Squared increases with the predictors
  - Even true for irrelevant predictors

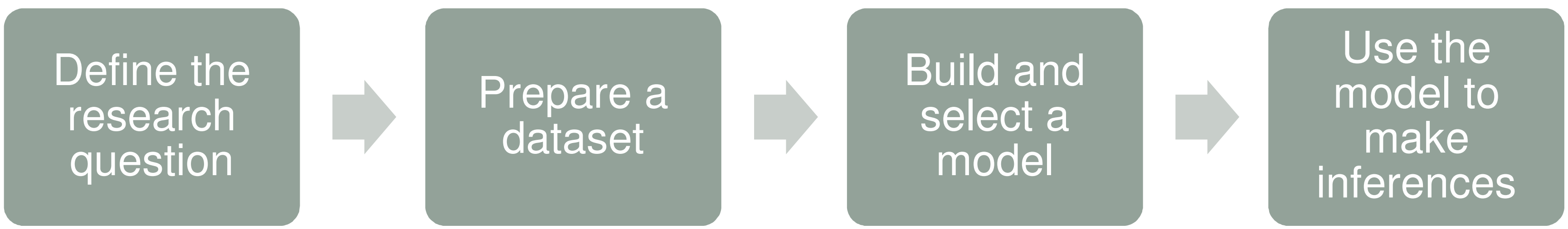
- When we have a predictor for each data point, R-Squared is 100%

- Need to use metrics that allow us to prefer regressions where all parameters are relevant

Likelihood

- The most important concept in model fitting
- Essentially writes the probability of observing the data given the model
- Most models are fit by finding a parameterization that maximizes the likelihood
  - Technical term is “Maximum Likelihood Estimation”
- We’ve actually been using likelihoods all along
  - It’s even how we estimate simple percentages

Data Analysis Process



Likelihood - Example

- Say we have a coin that when flipped gives heads 60% of the time

- Suppose we observe the following sequence
  - HHTTHTHTH

- We calculate the likelihood by counting the number of heads and tails
  - Likelihood = .6^5 * .4^4
    = 0.00199

Log-Likelihood

- Calculating a likelihood can get difficult for large datasets
- This is just because the numbers get so small
- To avoid this we often take the log of the likelihood – or Log likelihood
- Since log is an increasing function – the value that maximizes the log likelihood also maximizes the likelihood
- Also makes math much easier (for those into Calculus)

Log-Likelihood

- Does the log likelihood account for the number of predictors?
- No….
- But it’s used in metrics that do

Explanatory Metrics

- Can figure out which model we should use out of a set of models
- Adjusted R-Squared
- AIC (Akaike Information Criterion)
- BIC (Bayesian Information Criterion)

Adjusted R-squared

- The adjusted R^2 is a measure of explanatory power which is adjusted for the number of explanatory variables included in the regression.

- Adj-R^2 = 1 – (1-R^2) * (n-1)/(n-m-1),

where n is #observations, m is #explanatory variables.

AIC

- Stands for Akaike Information Criterion
- AIC = 2*k – 2*logLikelihood
- k is the no. of parameters in the statistical model.
- L is the maximized value of the likelihood function of the estimated model.
- When comparing a set of models for the data, the preferred model is the one with the lowest value of AIC.

BIC

- Stands for Bayes Information Criterion
- BIC = k*ln(n) - 2*logLikelihood
- n is the number of observations (or data points).
- k is the number of parameters.
- When comparing a set of models for the data, the preferred model is the one with the lowest value of BIC.
- In general, the penalty term is larger in BIC than in AIC.

Explanatory Metrics – How To Get From R

- In a Regression, all of these can be pulled from a regression
- Adjusted R-Squared is pulled from summary
- AIC and BIC have their own functions

```
y = c(1,5,3,10)
x = c(1,2,3, 4)
basicLM = lm(y~x)
#Find adjusted R-Squared in the summary
command
names(summary(basicLM))
AIC(basicLM)
BIC(basicLM)
```

Likelihood Ratio Tests

- These are all good, but aren’t formal statistical tests
  - We can’t assign p-values
- Likelihood Ratio Test can but only for ‘nested’ models
  - i.e. it can compare two models when one model is a more general version of the other
- Because one model is more general, it will have a higher likelihood
- The Likelihood Ratio Test computes how much higher this is

Likelihood Ratio Test (LRT)

- Let *Lr* be the maximized likelihood of a restricted model with *Pr* parameters
- Let *Lu* be the maximized likelihood of the unrestricted model with *Pu*
- Test Statistic: $-2[\ln(Lr) - \ln(Lu)]$
- Null hypothesis is that this has a chi-squared distribution with *Pu* – *Pr* degrees of freedom

Quick Take: Segmentation and Factor Analysis

- This is covered in depth in market research
- Both of these can be considered an explanatory analysis
- Data Reduction Processes
    - Instead of thinking about 1000 consumers, we think of 4 representative ones
- Keep in mind that these are required for humans, not computers

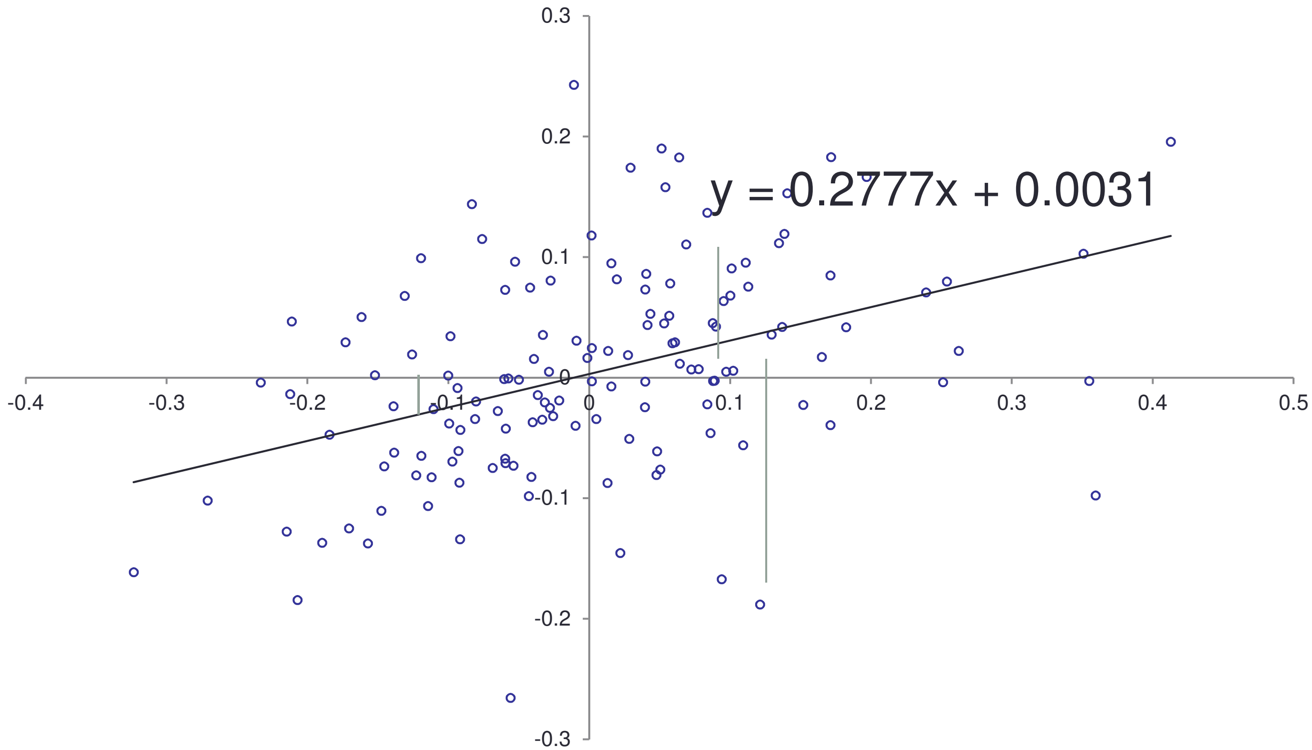
FORECASTING

Forecasting - Agenda

- These handle time series data
- Linear Regression Model
- Autoregressive Process (AR-Models)
- Moving Average Process (MA-Models)
- Need to have these two concepts down!

Forecasting - Agenda

- If you do understand those two, the following topics will be much easier
  - ARMA models
  - ARCH models
  - GARCH models
- Lots of topics, so we’ll review what we’ve learned a few times



y = 0.2777x + 0.0031
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5

Four Regression Assumptions

1. Linearity
   - The relationship between the independent variable and the dependent variable is best-approximated with a straight line
2. Independence
   - The errors need to be unrelated to each other
   - As usual, we can't prove this to be true. We need to rely on our understanding of the data generating process.
3. Equal variance
   - In order to conduct inference, the errors need to be drawn from the same distribution
4. Normality
   - In small samples, normal errors make calculating the t-statistic relevant

Time Series Data

- Time Series Data is simply data where we have multiple observations over time
- We can use the ‘lag’ function with this kind of data
  - This just means ‘take the observation in the previous period’
- We’ve used time series data in assignment 1 and 2
  - i.e. Sales of Peanut Butter over time

How do you feel today?

- Suppose at the end of the day, people asked you on a scale from 1 to 10 how you felt that day
- Our basic model is:

$$Feeling_t = News_t + e_t$$

- Regressions assume we have the correct functional form, and that the errors are uncorrelated and identically distributed

How do you feel today?

- If you felt good yesterday, you might feel good today
- In terms of regression, this could be represented as
  - $Feeling_t \sim News_t + Feeling_{t-1} + e_t$
- This is called an ***Autoregressive*** model

Autoregressive Model

- "Auto" means "Self" in Latin
  - An "autobiography" is a book where the writer writes about themselves
  - Also called an 'AR' model for short
- "Autoregressive" means we regress y on itself

$$Feeling_t \sim News_t + Feeling_{t-1} + e_t$$

- This is an auto-regressive model of degree-1

Autoregressive Model - Degree

- Degree simply references how many ‘lags’ we are controlling for

$$Feeling_t \sim News_t + Feeling_{t-1} + Feeling_{t-2} + e_t$$

- This is model has 2 lags, and so it is an auto-regressive model of degree-2
- AKA an AR-2 process

Moving Average Model

- Suppose you went out drinking last night
- Last night was awesome! I’ll give it a 9!
- This morning wasn’t as awesome. I’ll give a 4
- Not all 9’s are created equal!

Deciding what to do on the weekend

Went Out Drinking

- Rating for the day: 8
- Rating for the next day: 4

- Just controlling for the previous
day's rating might not be enough
- Not all '8's are created equal!

Did Homework

- Rating for the night: 8
- Rating for the next day: 8

Error Term

- To account for this behavior, we need something in the regression that controls for the long-term effect of things we don’t observe
- Remember causality – where are those things stored?
- Anything we don’t model will end up in our error term
- So, to control for these things, we control for the previous error term

Moving Average Model

- Called a ‘moving average’ model being overall residuals is a moving average of other error terms
- For example, this is a Moving Average regression model
    - $Feeling_t \sim News_t + e_t + e_{t-1}$
- Note how there are two error terms, one from this period and one from the previous period

Moving Average Model

- Consider the 'drinking' example
- In one period, our *e* was high because we went out and had fun
- In the next period, we were hung over
  - How we feel in this period might be negatively related to our *e* in the previous period
- So our model is

$$\text{Feeling}_t \sim \text{News}_t + e_t + e_{t-1}$$

Moving Average Model

- “Moving Average” Models are also called “MA Models”

$$Feeling_t \sim News_t + e_t + e_{t-1}$$

- Moving average models also have degrees
- The above model is an ‘MA-Model with Degree-1’
- Below is a model with degree 2

$$Feeling_t \sim News_t + e_t + e_{t-1}$$

Autoregressive versus Moving Average

- Very frequently confused!
- Some ways of describing the difference:
  - "Autoregressive models" control for things we observe – our y
  - Hence the 'auto' name – we regress *y* on itself
  - "Moving Average" models control for things we don't observe – our *e*

Regression doesn’t naturally handle this

- Both these types of terms may be omitted variables that can cause bias
- Furthermore, one assumption in a basic linear regression is that residuals are independent over time
- In the case of time series, residuals may be correlated over time
- This is exactly what an MA process handles

ARMA-Models

- We’ve done AR models
- We’ve done MA models
- Now we will do ARMA models
- Any guesses as to what these are?

ARMA Models

- ARMA Models combine AR and MA models
- The ***autoregressive*** component includes previous *y*
- The ***moving average*** component includes previous error terms *e*
- An ARMA model might look like this:

$$\text{Feeling}_t \sim \text{Feeling}_{t-1} + \text{Feeling}_{t-2} + e_t + e_{t-1}$$

ARMA Models - Degree

- Each of these can have a different degree degree
- An 'ARMA(p,q)' model includes an AR process with degree *p* and an MA process of degree *q*
- What is the degree of the models below?

$$Feeling_t \sim Feeling_t + Feeling_{t-1} + Feeling_{t-2} + e_t + e_{t-1}$$

$$Feeling_t \sim Feeling_t + Feeling_{t-1} + Feeling_{t-2} + Feeling_{t-3} + e_t + e_{t-1} + e_{t-2}$$

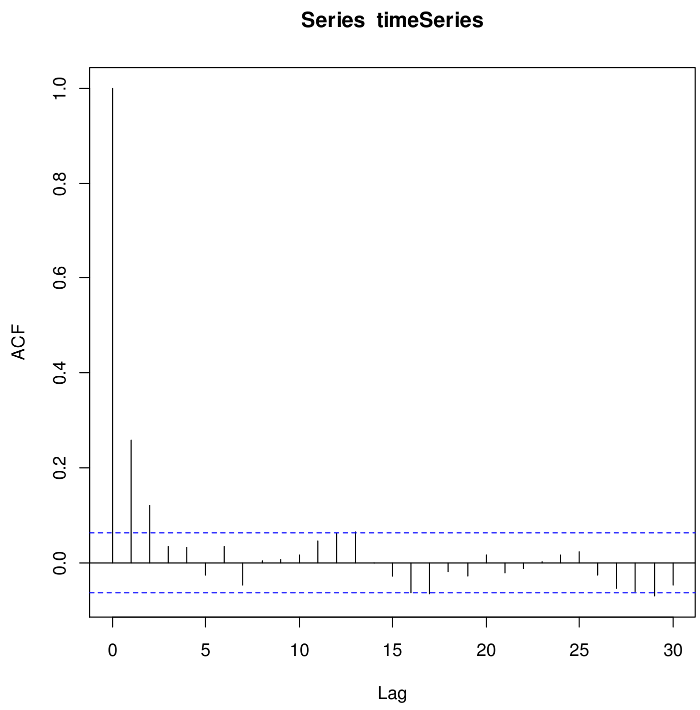
ARMA Models

- Many possible models, how do you pick the right one?
- **Classical Statistics**: Look at diagnostic plots and know what they mean
- **Modern Statistics**: You tell me – how might we evaluate a model like this?

Autocorrelation Plots

- One way to find out the ARMA model you need
- Computes how correlated the *y* are over time
- Goes through a series of 'lags' and calculates the correlation
- In general, it shows the correlation between y_t and y_{t-k} while changing *k*
- A simple linear regression would list all these correlations as 0

Sample Autocorrelation Plot



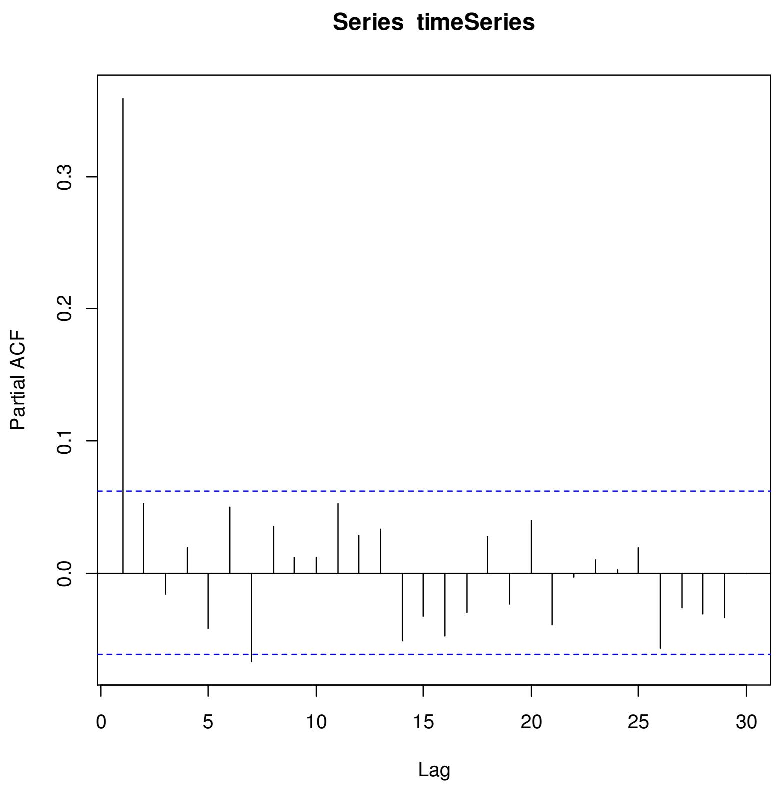
Autocorrelation Plot

- Suppose the true data comes from an AR(1) process
  - Aside – what does this model look like?
- Will y_t be correlated with y_{t-1}?
- Will y_{t-1} be correlated with y_{t-2}?
- Will y_t be correlated with y_{t-2}?
- Can’t tell from this

Partial Autocorrelation Plot

- This asks ‘What is the correlation between y_t and y_{t-k} controlling for the y in the middle
- Let’s say k = 2
- Then the partial auto correlation is the correlation between y_t and y_{t-2} controlling for y_{t-1}
- Essentially it runs a regression with $y_t \sim y_{t-1} + y_{t-2}$ and checks if the coefficient of y_{t-2} is significant
- If k = 3 – what is the partial autocorrelation?

Partial Autocorrelation Plot



Let's Recap what we've done so far

- We have two types of processes modelled
  - ***Autoregressive***, where the value today depends on previous values
  - ***Moving Average***, where the value today depends on previous errors

$$Feeling_t \sim \underbrace{Feeling_{t-1} + Feeling_{t-2} + Feeling_{t-3}}_{\text{Autoregressive Component}} + \underbrace{e_t + e_{t-1} + e_{t-2}}_{\text{Moving Average Component}}$$

- The ***Autocorrelation*** function measures the correlation between Y_t and Y_{t-k}
- The ***Partial Autocorrelation*** function measures the correlation between Y_t and Y_{t-k}, **controlling** for other correlations

What does an AR processes look like?

- Consider an AR-Process with degree 1
- Then we have the following equation:

$$Y_t = Y_{t-1} + e_t$$

- This implies

$$Y_5 = Y_4 + e_5$$

$$Y_4 = Y_3 + e_5$$

What does an AR process look like?

- This implies

$$Y_5 = Y_4 + e_5$$

$$Y_4 = Y_3 + e_5$$

- Is Y_5 correlated with Y_4?
- Is Y_5 correlated with Y_3? Sub-in for Y_4 to find out
- Which of these correlations should be stronger?
- What are the partial autocorrelations? Between Y_5, Y_4, and Y_3?

What does an MA processes look like?

- Consider an MA-Process with degree 1
- Then we have the following equation:

$$Y_t = e_t + e_{t-1}$$

- This implies

$$Y_5 = e_5 + e_4$$
$$Y_4 = e_4 + e_3$$
$$Y_3 = e_3 + e_2$$

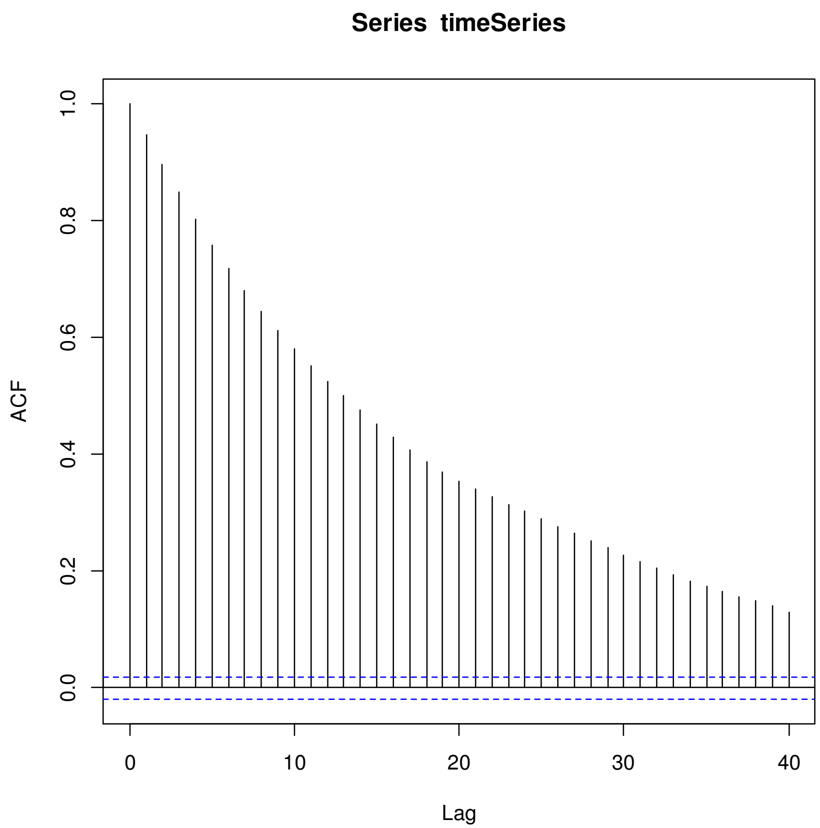
- Is Y_5 correlated with Y_4?
- Is Y_5 correlated with Y_3?

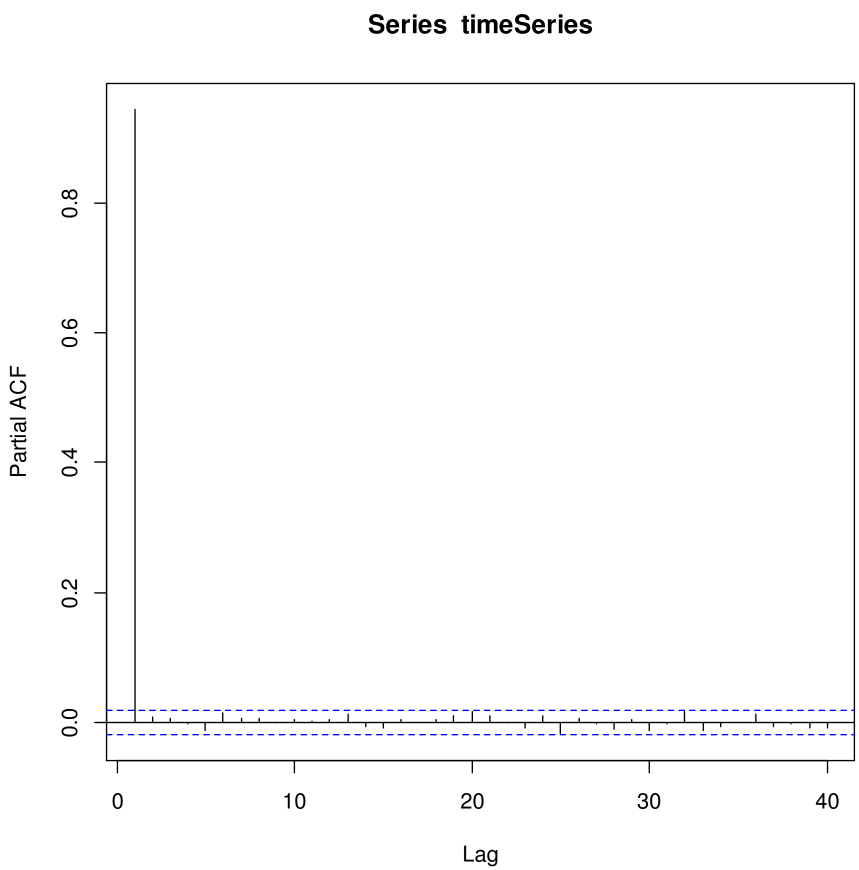
What do AR and MA processes look like?

- In an MA process, there will be autocorrelation between Y_t and a fixed number of lags
  - i.e. In an MA-Degree 1, there will be autocorrelation ONLY with Y_{t-1}
- In an AR process, the auto correlation will decline with the lag
  - I.e. In an AR-Degree 1, Y_t will be autocorrelated with Y_{t-1}, Y_{t-2} etc.
  - But the magnitude will decline
- The degree of the AR process can be found using the partial autocorrelation

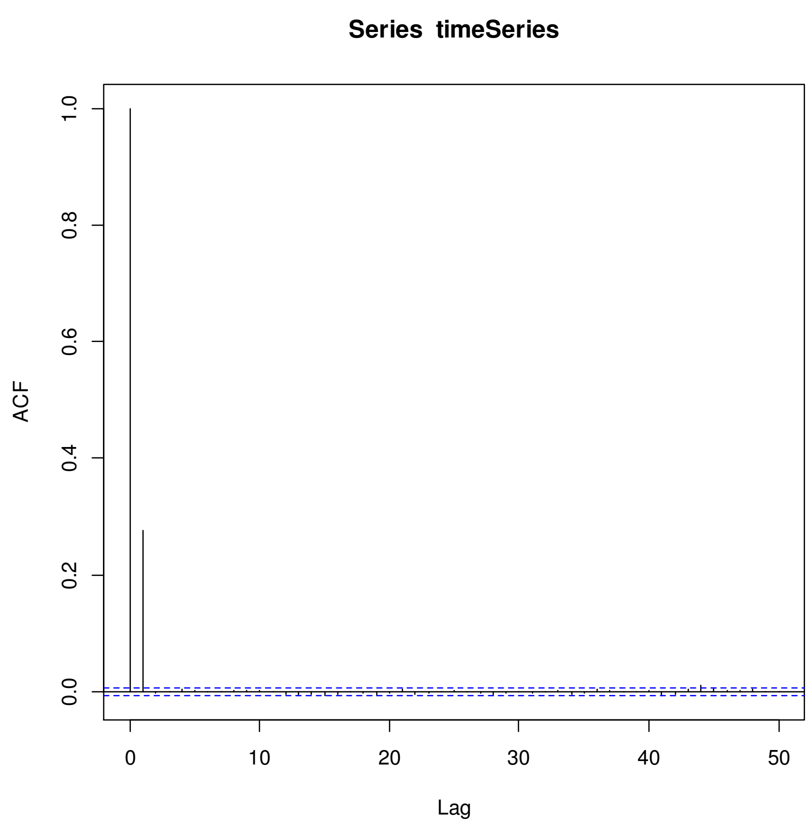
Shape	Indicated Model
Exponential, decaying to zero	Autoregressive model. Use the partial autocorrelation plot to identify the order of the autoregressive model.
Alternating positive and negative, decaying to zero	Autoregressive model. Use the partial autocorrelation plot to help identify the order.
One or more spikes, rest are essentially zero	Moving average model, order identified by where plot becomes zero.
Decay, starting after a few lags	Mixed autoregressive and moving average (ARMA) model.
All zero or close to zero	Data are essentially random.
High values at fixed intervals	Include seasonal autoregressive term.
No decay to zero	Series is not stationary.

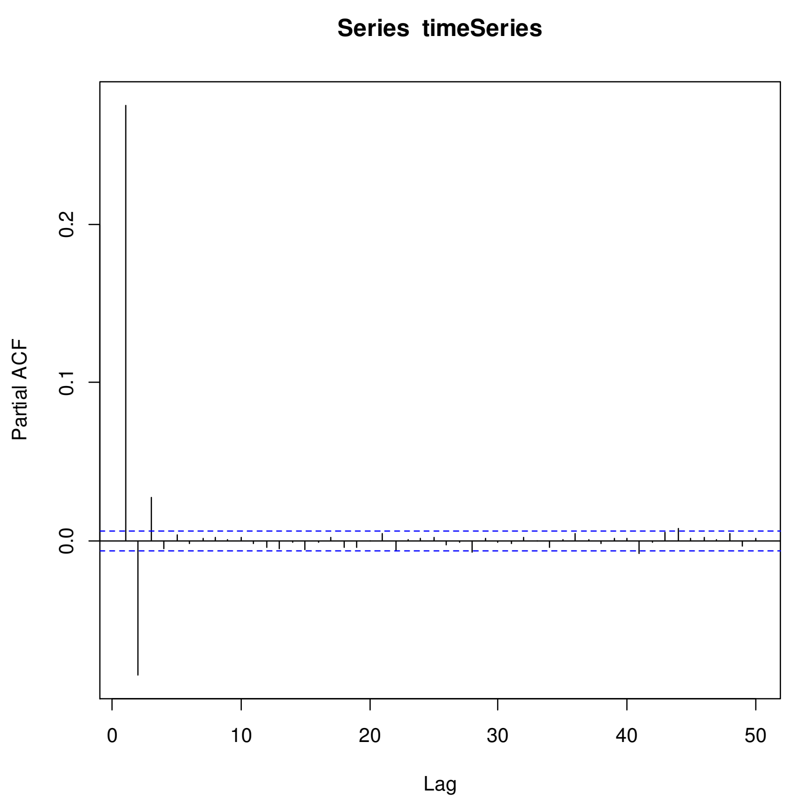
Let’s identify some ARMA processes





Let's identify some ARMA processes





ARMA Model - Code

```
fit1 <- arima(presidents, c(1, 0, 0))
fit2 <- arima(presidents, c(0, 0, 1))
fit3 <- arima(presidents, c(1, 0, 1))
fit4 <- arima(presidents, c(2, 0, 1))
```

Remember the other approach!

- Just run a bunch of models and take the one with the lowest AIC
- Much easier, possibly more accurate
- Less easy/impressive to explain to others

ARMA Model - Code

```
AIC(fit1)
AIC(fit2)
AIC(fit3)
AIC(fit4)
```

ARMA Model - Code

```
modelStat = data.frame(ar = rep(NA,25),ma =
rep(NA,25),AIC = rep(NA,25))
rowNum = 1
for(arDegree in 0:4){
     for(maDegree in 0:4){
          currentFit = arima(presidents,
c(arDegree, 0, maDegree))
          modelStat[rowNum,] =
c(arDegree,maDegree,AIC(currentFit))
          rowNum = rowNum + 1
     }
}
```

Seasonality

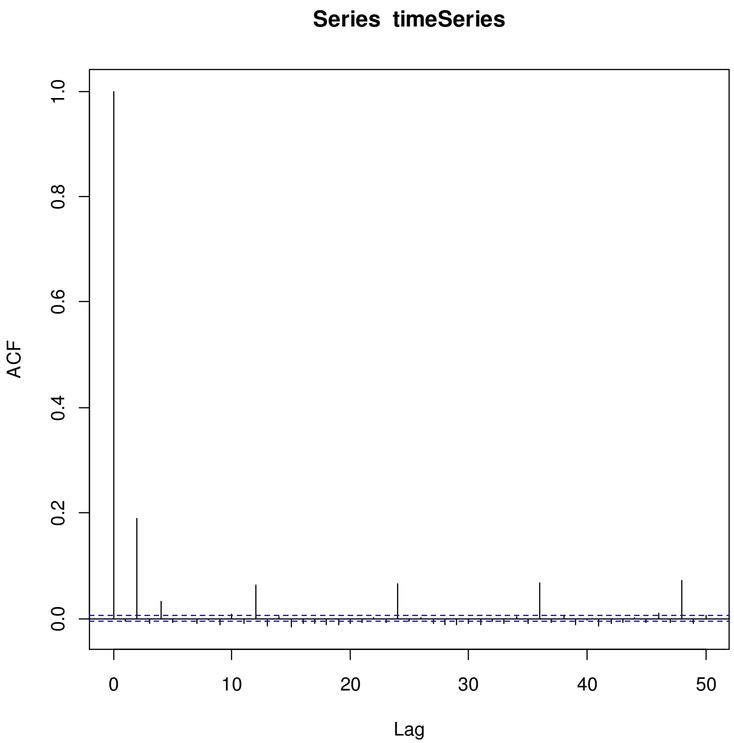
- Seasonality is tied to time series data
- You can use auto-correlation plots to check for seasonality
- Suppose we have monthly data, and we are investigating the sales of gazpacho
  - Gazpacho is a summer soup
- We initially run a model without seasonality

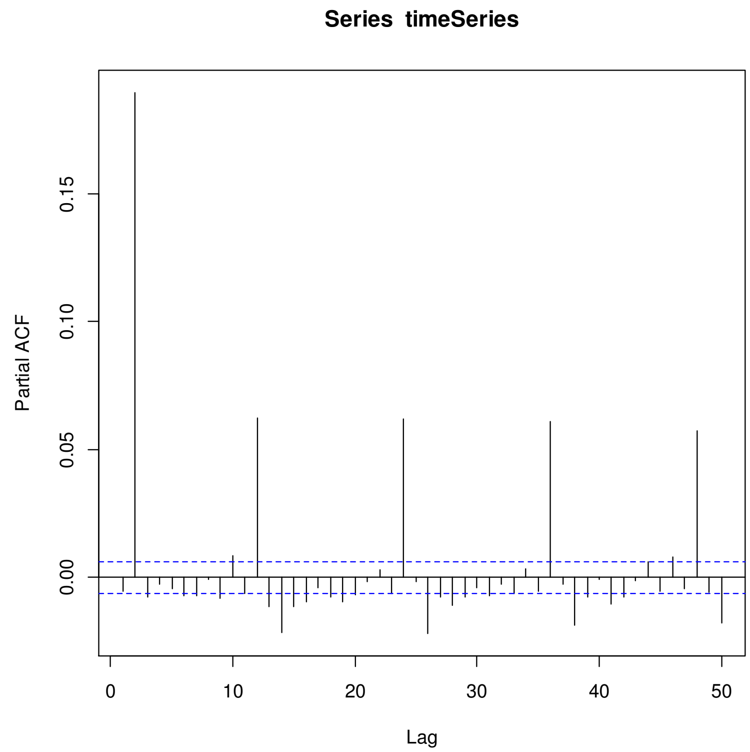
$$Sales_t = Price_t + e_t$$

Seasonality

- Since seasonality is not in the model, it *is* in the error term
- The error term will be higher in the summer, and lower in the winter
- With monthly data, we should see a positive correlation between months with a 12-month gap
  - The error terms in June 2013 and June 2014 are higher due to seasonality
  - The error terms in December 2013 and December 2014 are higher due to seasonality
- We should see a large, positive autocorrelation at a lag of 12

Seasonality ACF





Quick Quiz

- What additional terms get controlled for in an AR model?
- What additional terms get controlled for in an MA model?
- What trend do we look for in the auto correlation function for an AR model?
- What trend do we look for in the auto correlation function for an MA model?
- How do we see seasonality in an auto-correlation plot?

ARCH Models

- So far we have focused on the expectation of Y
- However, the variance of Y might also change over time
- Linear regression assumes that the variance of the error term is a fixed constant over time and across individuals
- If this assumption is violated, our standard errors could be wrong

ARCH Models

- ARCH stands for Auto-Regressive Conditional Heteroskedasticity
- Auto-Regressive: It depends on itself
- Conditional Heteroskedasticity: The variance is changing
- In an ARCH model, the **variance** follows an AR process
- That is, if e_t is drawn from a normal distribution with mean 0 and variance σ_t

ARCH Models

- In an ARCH model, the **variance** follows an AR process
- That is, if e_t is drawn from a normal distribution with mean 0 and variance σ_t
- In a standard linear regression, or an ARMA model,

$$\sigma_t = \sigma \text{ for all } t$$

- In an ARCH model with degree 1, variance is
  - $\sigma_t = \sigma_{t-1} + e_t'$
- Note that this e is different than the one in the regression model

ARCH Models

- This does well in cases where the variance seems to be changing over time
- Sudden Volatility Spikes
- More of a finance topic, but good to be aware of
- ARCH models are AR processes applied to the variance of the model

GARCH Models

- The G stands for "generalized"
- ARCH models are AR processes applied to the variance of the model
- GARCH models are ARMA processes applied to the variance of the model
- That is, in a GARCH model we can add a "Moving Average" component to the variance process

$$\sigma_t = \sigma_{t-1} + e_t' + e_{t-1}'$$

GARCH Models – When To Use?

- Several Approaches
- Simplest is again to fit a bunch and check the AIC
- Alternative approach is to run auto-correlation plots on the squared residuals, but this is not covered here
  - See wiki article for more details

Let's Recap what we've done so far

- We have two types of processes modelled
  - ***Autoregressive***, where the value today depends on previous values
  - ***Moving Average***, where the value today depends on previous errors

$$\text{Feeling}_t \sim \underbrace{\text{Feeling}_{t-1} + \text{Feeling}_{t-2} + \text{Feeling}_{t-3}}_{\text{Autoregressive Component}} + \underbrace{e_t + e_{t-1} + e_{t-2}}_{\text{Moving Average Component}}$$

- An ***ARMA*** model has both autoregressive and moving average components
- A ***GARCH*** model applies the ARMA framework to the variance

Panel Data

- Suppose we have multiple time series
  - Individual shoppers making purchases over time
  - Stores reporting their revenue over time
- Stores might have different levels of popularity
- Without accounting for individual differences, we might get the wrong model

Panel Data

- Two approaches to account for individual effects that can be combined
- Add individual “Fixed Effects” – just a factor variable
- Difference the data

Panel Data - Differencing

- Suppose we have the following model

$$Y_t \sim Y_{t-1} + \text{factor(storeEffect)} + e_t$$

- This implies that

$$Y_5 \sim Y_4 + \text{factor(storeEffect)} + e_5$$

$$Y_4 \sim Y_3 + \text{factor(storeEffect)} + e_4$$

- We can get rid of the store effect by ***differencing*** – subtracting one equation from another

$$Y_5 - Y_4 \sim Y_4 - Y_3 + e_5 - e_4$$

Panel Data

- Differencing can be implemented with the *arima* function in R
- Just change the second argument to '1'
- Can combine differencing and fixed effects
- That would capture that some stores are growing, and some are shrinking

Wrap Up