GENERAL ASSEMBLY

INTRODUCTION TO FORECASTING: ARIMA AND SEASONALITY

Joseph Nelson, Data Science Immersive

AGENDA

- ARIMA Review
- Overview: what are the steps to building an effective time series model?
- Dickey-Fuller Test
- ACF and PACF
- Ljung-Box Test
- Akaike Information Criteria
- Code Along

FRAMING: FORECASTING IS VERY HARD

- Forecasting models often take **months** to effectively tune. (Why do you think well-paid PhD economists are wrong so often yet still employed?!)
- Today's lecture is meant to be an introduction to how statisticians/data scientists approach modeling time series + a practical walkthrough of an example.

FRAMING: FORECASTING IS VERY HARD

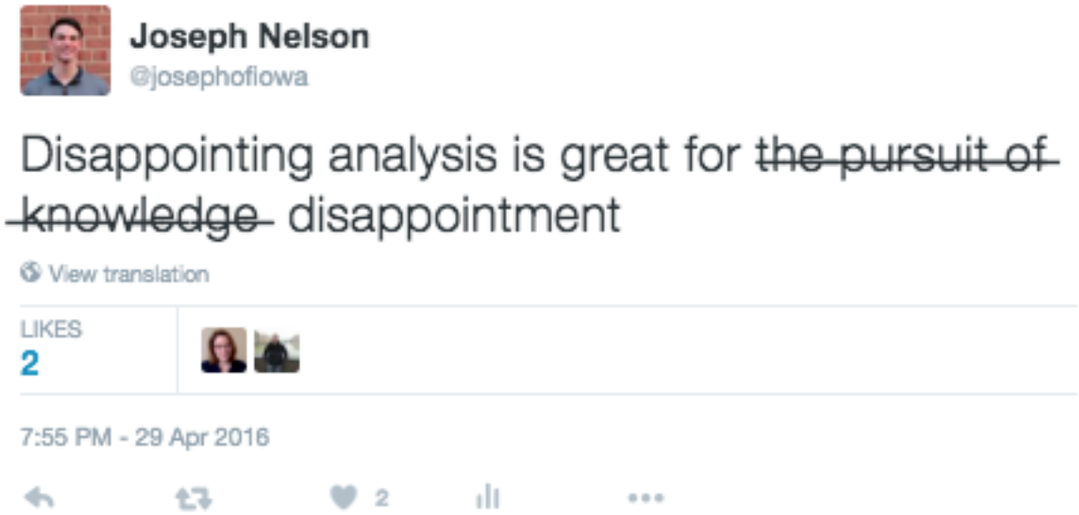
- To prove this point…Some self-deprecating humor

- So, remember:

FRAMING: FORECASTING IS VERY HARD

- To prove this point…Some self-deprecating humor

- So, remember:



REVIEW: ARIMA

AR + MA

- **Autoregressive (AR)** models are those are that use data from previous time-points to predict the next time-point. These are very similar to previous regression models, except as input - we'll take some previous outcome.
- We must select some value, p, for the amount of lag we believe to be useful for predicting future values.
- How might we identify what p value we should use?

REVIEW: ARIMA

AR + MA

- **Autoregressive (AR)** models are those are that use data from previous time-points to predict the next time-point. These are very similar to previous regression models, except as input - we'll take some previous outcome.
- We must select some value, p, for the amount of lag we believe to be useful for predicting future values.
- **Moving average** models, as opposed to autoregressive models, do not take the previous outputs (or values) as inputs, but instead take the previous error terms. We will attempt to predict the next value based on the overall average and how incorrect our previous predictions were.
- We must select some value, q, for the number of previous errors we consider.
- What does a MA of prior error help capture in our model, in particular?

ARIMA

- What does the “i” stand for? Why does that matter?

ARIMA

- What does the “i” stand for? Why does that matter?
- “i” stands for integrated: we are combining AR and MA techniques into a single model: ARIMA.
- Integrating the two tactics results in us selecting some differencing term, d, where we are now predicting the DIFFERENCE between one prior period and the new period, rather than predicting the new period’s value itself.
- ie: $y_t - y_{(t-1)} = ARMA(p, q)$
- This is important because it helps us de-trend our data and approach stationarity.

ARIMA

- We now know there are three values we may consider tuning with a standard ARIMA model: AR(p) MA(q) and differencing (d)
- Recall: what did we say p may reflect? What about q? And d?

ARIMA(p, d, q)

- We now know there are three values we may consider tuning with a standard ARIMA model: AR(p) MA(q) and differencing (d)
- Recall: what did we say p may reflect? What about q? And d?
- Our p indicates how many prior time periods we're taking into consideration for explained autocorrelation. Increasing p would increase the dependency on previous values further (longer lag).
- Our q indicates how many prior time periods we're considering for observing sudden trend changes.
- Our d indicates what difference we are anticipating predicting. d=1 may cause stationarity for us; d=2 may capture exponential movements

OVERVIEW: TIME SERIES FORECASTING STEP-BY-STEP

1. Visualize the time series
2. Stationize the series
3. Plot ACF/PACF to Seek Optimal Parameters
4. Build the ARIMA Model
5. Predict the Future

"Park the bus"

OVERVIEW: TIME SERIES FORECASTING STEP-BY-STEP

1. Visualize the time series
2. Stationize the series

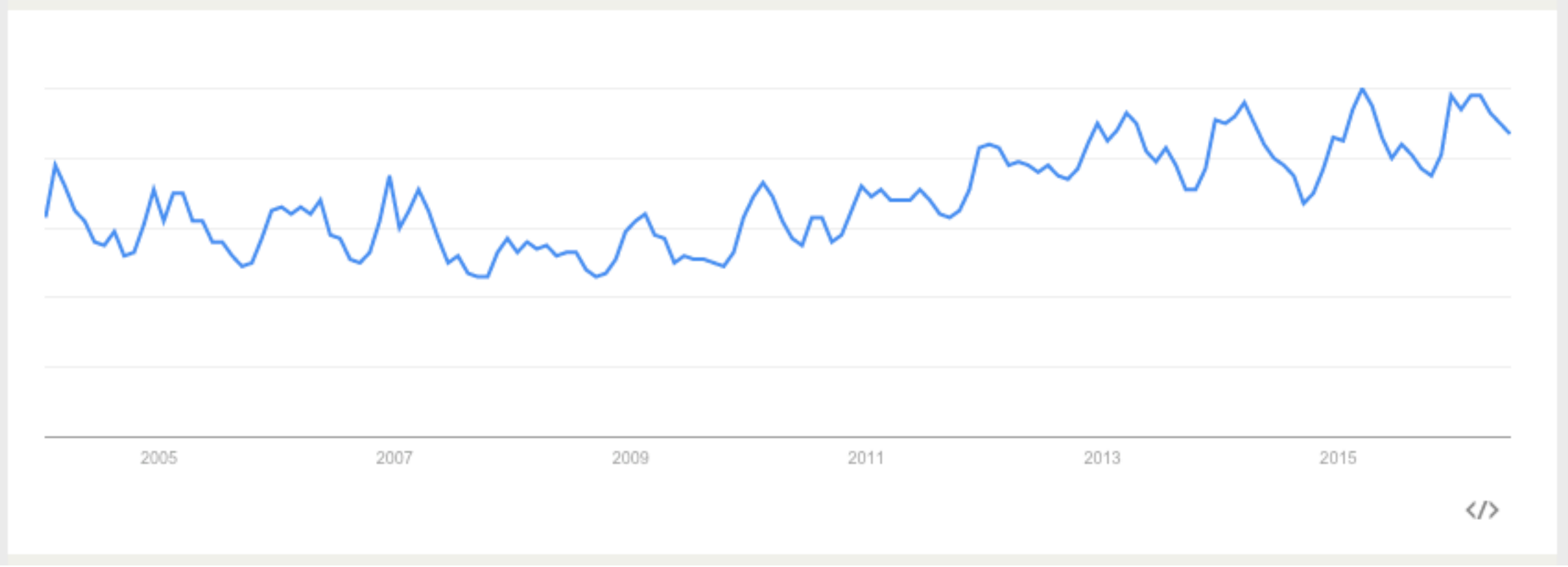
"Park the bus"

3. Plot ACF/PACF to Seek Optimal Parameters
4. Build the ARIMA Model
5. Predict the Future
6. Profit*

*Lecture sold separately

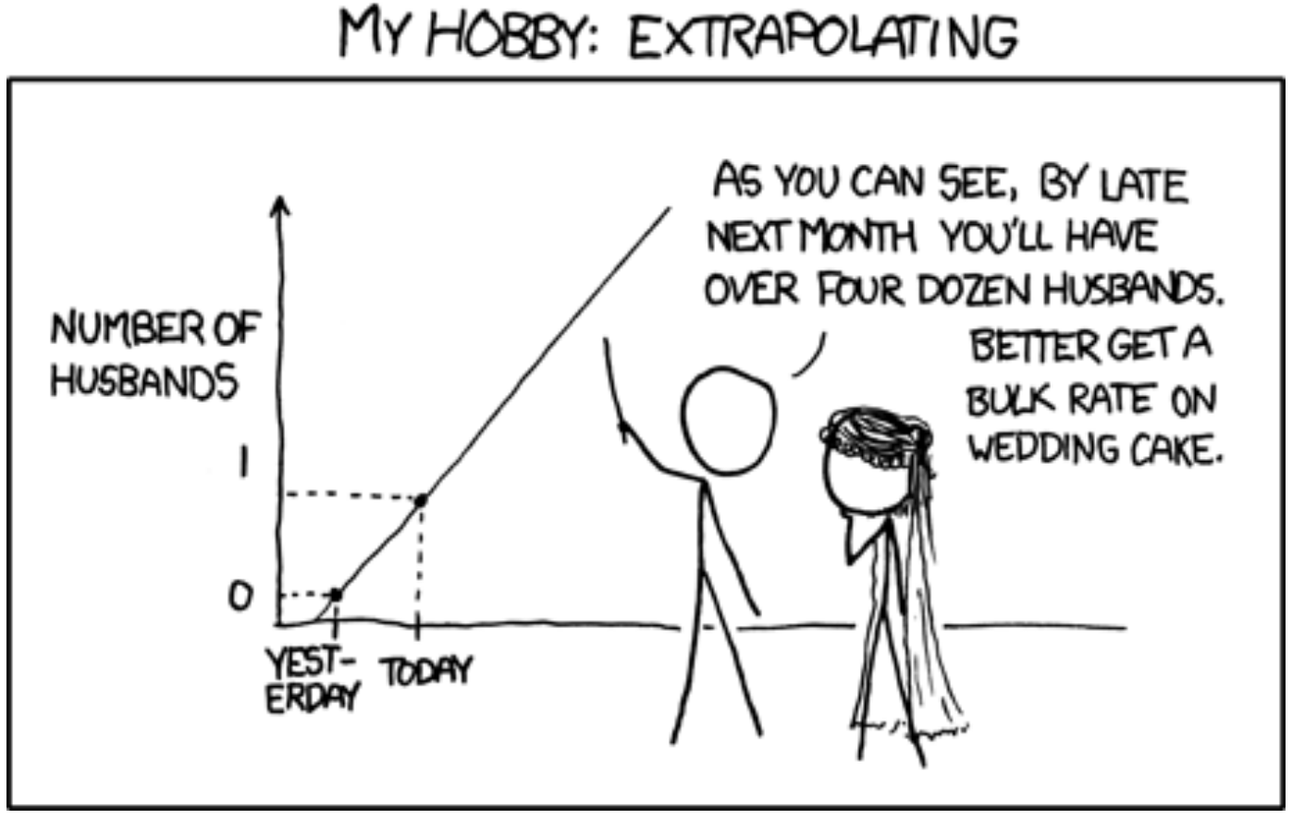
VISUALIZE THE DATA/EDA

- Is there a clear trend in the data?
- Does time seem explanatory?
- Are there regular periods/seasonality we observe?



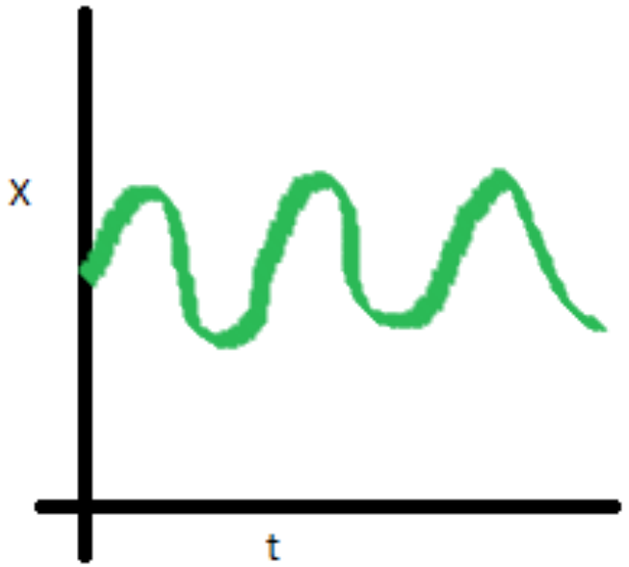
VISUALIZE THE DATA/EDA

- Is there a clear trend in the data?
- Does time seem explanatory?
- Are there regular periods/seasonality we observe?

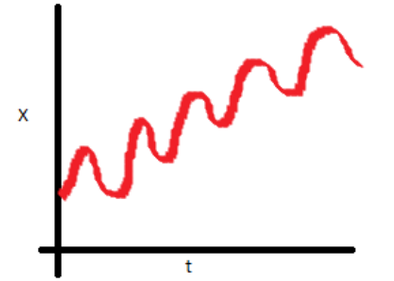


STATIONIZE THE DATA (“PARK THE BUS”)

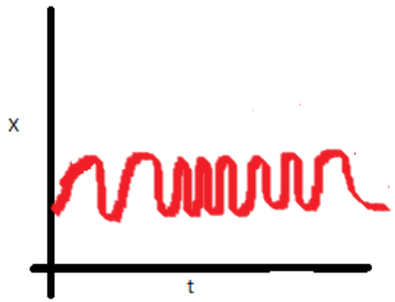
- Assure that the data can be viewed independent of time dependence. This enables us to apply a host of new analysis tactics to our dataset.



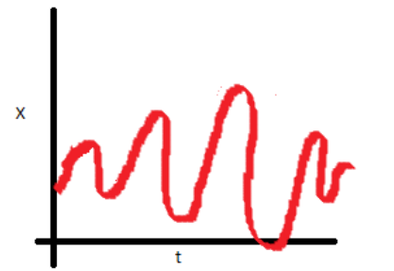
Stationary series



- 1



- 2



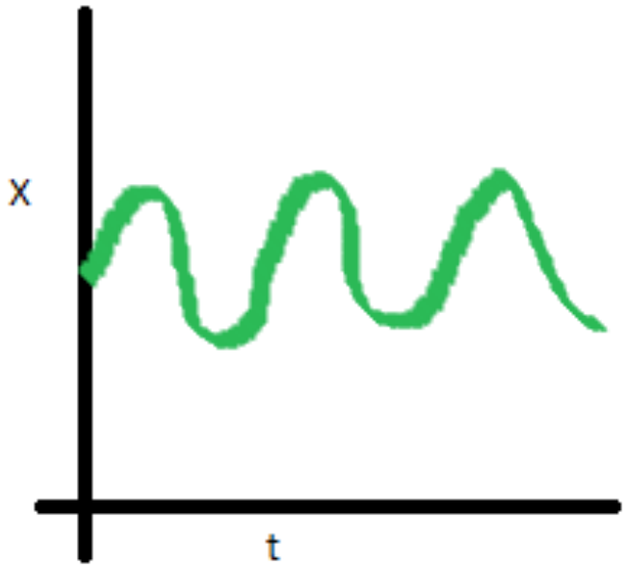
- 3

- **Match the non-stationary red series with the problem it exhibits on the right.**

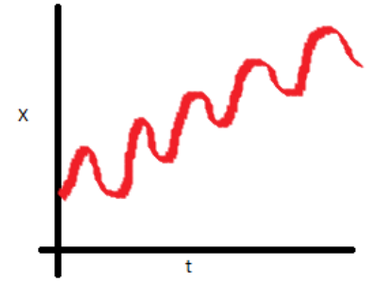
- A. Non-constant covariance
- B. Heteroskedastic
- C. Increasing mean

STATIONIZE THE DATA (“PARK THE BUS”)

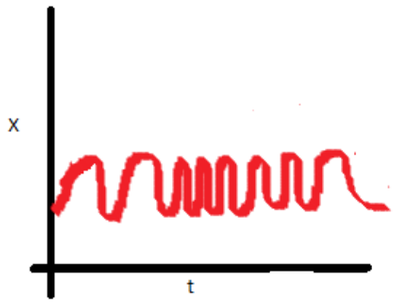
- Assure that the data can be viewed independent of time dependence. This enables us to apply a host of new analysis tactics to our dataset.



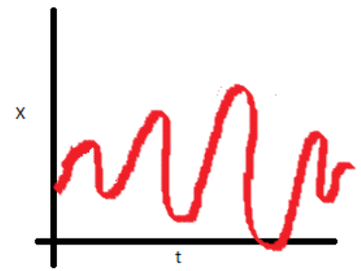
Stationary series



- 1: C



- 2: A



- 3: B

- **Match the non-stationary red series with the problem it exhibits on the right.**

  - A. Non-constant covariance
  - B. Heteroskedastic
  - C. Increasing mean

HOW DO WE ACHIEVE STATIONARITY?

- We can perform various transformations on our data in an effort to achieve stationarity.
- Examples include: deflation by CPI or other price index, deflation at a fixed rate, logarithm, first difference, seasonal difference, seasonal adjustment.
- Additional information on each of these tactics (including why you may choose a given tactic) is available here: http://people.duke.edu/~rnau/whatuse.htm

HOW DO WE ACHIEVE STATIONARITY?

- We can perform various transformations on our data in an effort to achieve stationarity.
- Examples include: deflation by CPI or other price index, deflation at a fixed rate, logarithm, first difference, seasonal difference, seasonal adjustment.
- Additional information on each of these tactics (including why you may choose a given tactic) is available here: http://people.duke.edu/~rnau/whatuse.htm

- Quiz: what tactic does the ARIMA model we previously discussed include?

HOW DO WE KNOW WE'RE STATIONARY? DICKEY-FULLER TEST

- 1.) Observing our data – check our transformed plots
- 2.) We can also use a statistical test to determine if our data is truly stationary

- For brevity, we'll discuss the test's output: if the 'Test Statistic' is greater than the 'Critical Value' than the time series is stationary.

- http://stats.stackexchange.com/questions/44647/which-dickey-fuller-test-should-i-apply-to-a-time-series-with-an-underlying-mode
- https://en.wikipedia.org/wiki/Dickey%E2%80%93Fuller_test

HOW DO WE SELECT GOOD MODEL PARAMETERS?

ARIMA(p, d, q)

- We know ARIMA enables us to identify regular autocorrelations (p), stationize our data (d), and anticipate trend shocks in our error terms (q)

- What would I be saying by passing values of (1,0,2)?

HOW DO WE SELECT GOOD MODEL PARAMETERS?

ARIMA(p, d, q)

- We know ARIMA enables us to identify regular autocorrelations (p), stationize our data (d), and anticipate trend shocks in our error terms (q)

- What would I be saying by passing values of (1,0,2)?
- (1,0,2) implies each term is correlated with one output prior (p=1). The two prior error terms may be useful in predicting my next output (q=2). My data is already stationary (d=0).

- How could we account for an error that occurs regularly, but not in every time period? (In other words, what if there are seasonal effects on my data?)

HOW DO WE SELECT GOOD MODEL PARAMETERS?

$$\text{ARIMA}\ \underbrace{(p,d,q)}_{\text{Non-seasonal part of the model}}\ \underbrace{(P,D,Q)_m}_{\text{Seasonal part of the model}}$$

- We can use Seasonal ARIMA to capture these effects. Based on our model:
- p = non-seasonal AR order
- d = non-seasonal differencing
- q = non-seasonal MA order
- P = seasonal AR order
- D = seasonal differencing
- Q = seasonal MA order m = number of periods per season

HOW DO WE SELECT GOOD MODEL PARAMETERS? ACF AND PACF

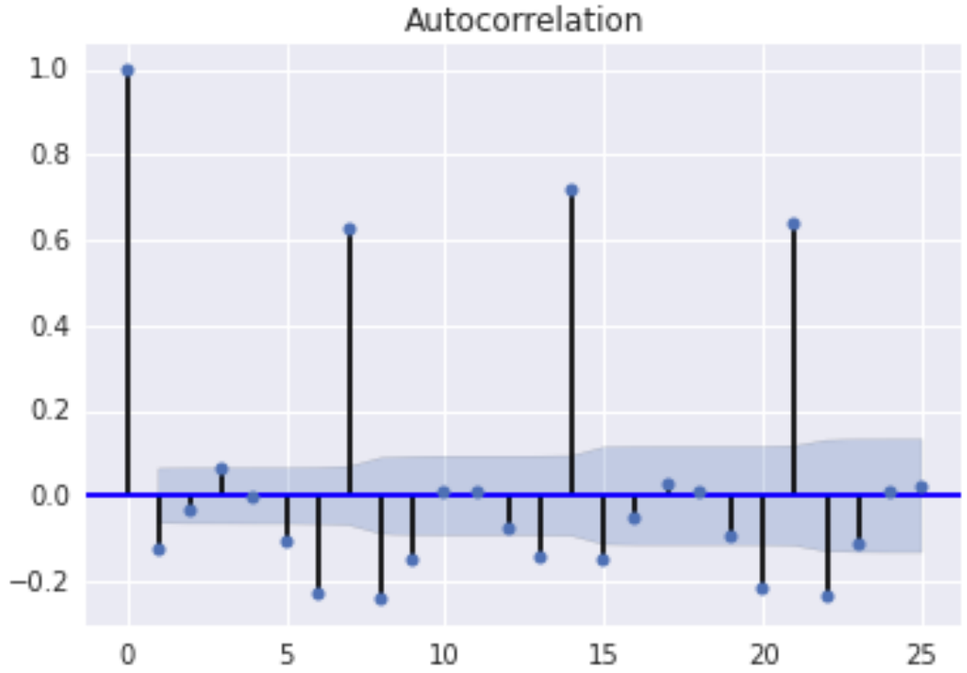
$$\text{ARIMA} \quad \underbrace{(p,d,q)}_{\substack{\uparrow \\ \text{Non-seasonal part} \\ \text{of the model}}} \quad \underbrace{(P,D,Q)_m}_{\substack{\uparrow \\ \text{Seasonal part} \\ \text{of the model}}}$$

- The ACF is the autocorrelated function: how much a variable is correlated with itself by lag time. The PACF is the partial correlation function by lag time.
- In general, the "partial" correlation between two variables is the amount of correlation between them which is not explained by their mutual correlations with a specified set of other variables. For example, if we are regressing a variable Y on other variables X1, X2, and X3, the partial correlation between Y and X3 is the amount of correlation between Y and X3 that is not explained by their common correlations with X1 and X2. This partial correlation can be computed as the square root of the reduction in variance that is achieved by adding X3 to the regression of Y on X1 and X2.

HOW DO WE SELECT GOOD MODEL PARAMETERS? ACF AND PACF

$$\text{ARIMA}\quad \underbrace{(p,d,q)}_{\substack{\uparrow\\ \text{Non-seasonal part}\\ \text{of the model}}}\quad \underbrace{(P,D,Q)_m}_{\substack{\uparrow\\ \text{Seasonal part}\\ \text{of the model}}}$$

- ACF for the German drug store sale data:



- What do you notice? What does this mean? How does it impact our model building?

HOW DO WE SELECT GOOD MODEL PARAMETERS? ACF AND PACF

$$\text{ARIMA}\quad \underbrace{(p,d,q)}_{\substack{\uparrow\\ \text{Non-seasonal part}\\ \text{of the model}}}\quad \underbrace{(P,D,Q)_m}_{\substack{\uparrow\\ \text{Seasonal part}\\ \text{of the model}}}$$

- What does the following imply about our ACF and PACF? ARIMA(0,0,0)(0,0,1)12

HOW DO WE SELECT GOOD MODEL PARAMETERS? ACF AND PACF

$$\text{ARIMA}\ \underbrace{(p,d,q)}_{\substack{\uparrow \\ \text{Non-seasonal part} \\ \text{of the model}}}\ \underbrace{(P,D,Q)_m}_{\substack{\uparrow \\ \text{Seasonal part} \\ \text{of the model}}}$$

- What does the following imply about our ACF and PACF? ARIMA(0,0,0)(0,0,1)12
- We are expecting seasonal intervals of 12 months. In our ACF, we may see a significant spike at k=12, but none others. We capture this **seasonal** 12 month spike's impact in our Q (the "random error getter"). Moreover, we would expect that the PACF demonstrates exponential decay at intervals of k=12—every future iteration of the 12th period is less powerful. It is a one-off event, and that is why we would consider it to be a part of Q.

HOW DO WE SELECT GOOD MODEL PARAMETERS? ACF AND PACF

$$\text{ARIMA}\ \underbrace{(p,d,q)}_{\substack{\uparrow \\ \text{Non-seasonal part} \\ \text{of the model}}}\ \underbrace{(P,D,Q)_m}_{\substack{\uparrow \\ \text{Seasonal part} \\ \text{of the model}}}$$

- What does the following imply about our ACF and PACF? ARIMA(0,0,0)(1,0,0)12
- We are expecting seasonal intervals of 12 months. In our PACF, we see a significant spike at 12, but no other significant spikes. That is, the k=12 point is adding a lot of signal to our model that we want to capture. At the same time, we would expect to see exponential decay in the ACF seasonal lags.
- Note: If the data you are investigating displays a seasonal pattern that is both strong and stable over time (i.e. temperatures are higher in the Summer, lower in the Winter), then it is likely that your Seasonal Difference term D should be set to 1. This will prevent the seasonal pattern from "dying out" in the long-term forecasts.

EVALUATING OUR MODEL

- It should be clear that forecasting model evaluation is especially tied to the data scientist's exploratory data analysis and familiarity with the subject. Choosing useful p,d,q values and adding seasonal effects is almost entirely a context-driven endeavor. In addition, there are a few key tactics we can explore:
- 1.) Plotting our residuals
- If we do not observe a pattern in our residual error terms, we have succeeded
- 2.) Ljung-Box Test
- We mathematically test the above assumption: are our resideuals

- 3.) Akaike Information Criteria

EVALUATING OUR MODEL

- It should be clear that forecasting model evaluation is especially tied to the data scientist's exploratory data analysis and familiarity with the subject. Choosing useful p,d,q values and adding seasonal effects is almost entirely a context-driven endeavor. In addition, there are a few key tactics we can explore:
- 1.) Plotting our residuals

$$Q_{\mathrm{BP}} = n \sum_{k=1}^{h} \hat{\rho}_k^2,$$

- 2.) Ljung-Box Test
- We mathematically test the above assumption: are our residuals autocorrelated our randomly distributed?
- Instead of testing randomness at each distinct lag, it tests the "overall" randomness based on a number of lags.
- The data scientist must determine what value, k, lag works (we have a formula)
- 3.) Akaike Information Criteria

EVALUATING OUR MODEL

- It should be clear that forecasting model evaluation is especially tied to the data scientist's exploratory data analysis and familiarity with the subject. Choosing useful p,d,q values and adding seasonal effects is almost entirely a context-driven endeavor. In addition, there are a few key tactics we can explore:
- 1.) Plotting our residuals
- 2.) Ljung-Box Test
- 3.) Akaike Information Criteria
- A **relative** measure of information gain from our model. Lower AIC values are better.
- It cannot tell us quality in an absolute sense.
- Parsimonious models are the goal—as few features (includes lags) as possible
- https://en.wikipedia.org/wiki/Akaike_information_criterion

CODE EXAMPLES

- To the repo…

- But wait, you'll want these:
- https://datamarket.com/data/set/22w6/portland-oregon-average-monthly-bus-ridership-100-january-1973-through-june-1982-n114#!ds=22w6&display=line OR https://raw.githubusercontent.com/ga-students/DSI-DC-1/master/week-09/4.1-ARIMA_pt2/portland-oregon-average-monthly-.csv?token=ANUtezW9Hk-YY182YnV8T1EZkyv9bvOuks5XbRXkwA%3D%3D
- What current version of statsmodels are you running?