Looking for Trends

300958 Social Web Analytics

WESTERN SYDNEY
UNIVERSITY

School of Computing, Engineering and Mathematics

Week 10

Outline

Looking for Trends

Looking for Trends in Facebook reach or Twitter mentions can relate to important business problems. For example,

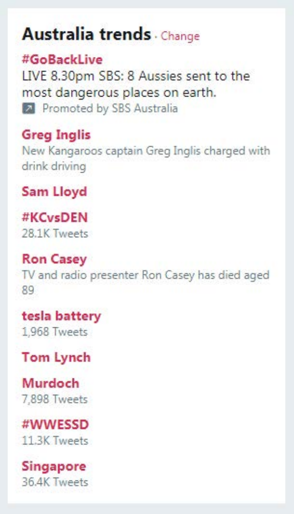
- Is our online profile increasing or decreasing.
- Are our competitors profiles increasing or decreasing.
- What are the trending topics? Twitter do this.
- Did some event have an impact? (Advertising campaign, product release?)

Focusing on Twitter, the first two relate to looking at the number of mentions over time. The third, Twitter already do, and we will look at how. The fourth, event impact, will be looked at in another lecture.

Outline

Twitter Trending Topics

Twitter produce a list of Trending topics. These can be made location specific e.g. Australian trends. But the basic technique for determining them is the same.
Taken from a presentation by Kostas Tsioutsioulikis.



What is a Topic?

Firstly, though we need to decide what is a topic?

- Can be a simple commonly occurring word or phrase
  - Of course, we would ignore stop words etc.
  - There is a problem, because there are a large number of possibilities.
- Another possibility is to use a dictionary.
  - Extract common phrases from other sources e.g. Wikipedia, usernames, sources specific to an application.

Given a (large) list of topics, what is trending?

What is Trending? Simple counting.

- Count the number of times a topic occurs (in Australia?) in a fixed time period.
- Look at the ratio of current frequency to the past.
- High ratios, mean a trend

Problems

- What about a new topic? Past frequency will zero (or close)
- What about overall frequency? Is an increase from 10 to 20 the same as 10,000 to 20,000?
- What is a high ratio?

Trends using a χ^2-test

WESTERN SYDNEY UNIVERSITY

Alternative is to establish an expected frequency using historical data, and compare the observed frequency (current) to the expected frequency (past). We can compare observed to expected using a χ^2 statistic

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

For a chi-squared test we need to consider a fixed number of tweets and compare those containing various topics.

Topic	A	B	C
Num. tweets Current	201	323	241
Num. tweets Past	181	299	285

Trends using a χ^2-test

There are problems with this approach, not least, some tweets may contain multiple topics. So Twitter use a simplified version.

Num Tweets	Tweet contains topic	Tweet doesn't contain topic
Current	201	564
Past	181	584

(This is equivalent to Proportion test, but Twitter uses χ^2)

Twitter's χ^2 score

Procedure;

- collect N tweets from past, and N tweets from today. (In fact N needn't be the same)
- Count X, the number of past tweets that contain the topic of interest.
- Count Y, the number of current tweets that contain the topic of interest.
- Calculate χ^2 statistic (treats past as expected)

$$\begin{aligned}\chi^2 &= \frac{(Y-X)^2}{X} + \frac{((N-Y)-(N-X))^2}{(N-X)} \\ &= \frac{(Y-X)^2}{X} + \frac{(X-Y)^2}{(N-X)}\end{aligned}$$

If $N >> X, Y$ then the second term will be very small and can be ignored.

Trending grand final

WESTERN SYDNEY UNIVERSITY

Problem

Out of five hundred tweets, 17 mention topic "grand final" in a past period, and 27 in the current. Calculate Twitter's χ^2 score.

$$\chi^2 = \frac{(Y - X)^2}{X} + \frac{(X - Y)^2}{(N - X)}$$

Trending grand final

Problem

Out of five hundred tweets, 17 mention topic “grand final” in a past period, and 27 in the current. Calculate Twitter’s χ^2 score.

$$\chi^2 = \frac{(Y - X)^2}{X} + \frac{(X - Y)^2}{(N - X)}$$

Note that we get the same χ^2 value if $Y = 7$, so the χ^2 statistic is measuring change, not increase. We must ensure that $Y > X$ to detect a trending topic.

Trending grand final

Problem

Out of five hundred tweets, 17 mention topic "grand final" in a past period, and 27 in the current. Calculate Twitter's χ^2 score.

$$\chi^2 = \frac{(Y - X)^2}{X} + \frac{(X - Y)^2}{(N - X)}$$

Note that we get the same χ^2 value if $Y = 7$, so the χ^2 statistic is measuring change, not increase. We must ensure that $Y > X$ to detect a trending topic.

Warning

Both X and Y are a sample, meaning that we should expect variation in them even if there is no increasing or decreasing trend. So we should not be treating X as the expected frequency.

What should we be testing?

If there is no increasing or decreasing trend, then the proportion of tweets should not change over time, meaning that the tweet count should be independent of time:

- H_0: The proportion of on topic tweets is independent of time.
- H_A: The proportion of on topic tweets is not independent of time.

This is a χ^2 test for independence.

What should we be testing?

If there is no increasing or decreasing trend, then the proportion of tweets should not change over time, meaning that the tweet count should be independent of time:

- H_0: The proportion of on topic tweets is independent of time.
- H_A: The proportion of on topic tweets is not independent of time.

This is a χ^2 test for independence.

$$\chi^2 = \sum_i \sum_j \frac{(X_{ij} - np_i q_j)^2}{np_i q_j}$$

- X_{ij} as the count of the ith row, jth column,
- n as the sample size (sum of all counts),
- p_i as the expected proportion of ith row and
- q_j as the expected proportion of the jth column.
- $np_i q_j$ is the expected count in cell ij, assuming H_0.

Trending grand final again

Problem

Out of five hundred tweets, 17 mention topic ”grand final” in a past period, and 27 in the current. Calculate the χ^2 statistic when testing independence.

$$\chi^2 = \sum_i \sum_j \frac{(X_{ij} - np_i q_j)^2}{np_i q_j}$$

χ^2 distribution assuming H_0

We have shown how to compute the test statistic. If the test statistic is large, then we reject the Null hypothesis.

But how large does the χ^2 test statistic have to be to be large?

Let's examine the χ^2 distribution when H_0 is true (when both proportions are equal, meaning the difference in proportions is zero).

Trend randomisation distribution

- We have X out of N past tweets that contain the topic.
- We have Y out of N current tweets that contain the topic.
- If there is no trend, then we can swap past and current tweets without effecting the trend. Note that if there is a trend, this swapping will effect the trend, but we assume H_0 until we can reject it.

Trend randomisation distribution

- We have X out of N past tweets that contain the topic.
- We have Y out of N current tweets that contain the topic.
- If there is no trend, then we can swap past and current tweets without effecting the trend. Note that if there is a trend, this swapping will effect the trend, but we assume H_0 until we can reject it.

Randomisation process:

1. Combine the current and past tweets and randomly divide them into two groups of size N that represent past and current.
2. Using the two groups, we recompute the χ^2 statistic and store it.
3. Repeat at least 1000 times to obtain the χ^2 distribution for H_0.

Randomisation Example

If we have $N = 5, X = 1, Y = 4$, then we have the sample:

Past tweets					Current tweets				
1	0	0	0	0	1	1	1	1	0

Randomisation Example

If we have $N = 5, X = 1, Y = 4$, then we have the sample:

Past tweets					Current tweets				
1	0	0	0	0	1	1	1	1	0

We then randomly divide the 10 tweets into two groups of $N = 5$.

Past tweets					Current tweets				
1	1	0	1	0	0	1	0	1	0

Randomisation Example

If we have $N = 5, X = 1, Y = 4$, then we have the sample:

Past tweets					Current tweets				
1	0	0	0	0	1	1	1	1	0

We then randomly divide the 10 tweets into two groups of $N = 5$.

Past tweets					Current tweets				
1	1	0	1	0	0	1	0	1	0

$\rightarrow$

	Topic	No Topic
Past	3	2
Curr	2	3

Randomisation Example

If we have $N = 5, X = 1, Y = 4$, then we have the sample:

Past tweets					Current tweets				
1	0	0	0	0	1	1	1	1	0

We then randomly divide the 10 tweets into two groups of $N = 5$.

Past tweets					Current tweets				
1	1	0	1	0	0	1	0	1	0

$\rightarrow$

	Topic	No Topic
Past	3	2
Curr	2	3

Giving $\chi^2 = (3 - 10 \times 0.5 \times 0.5)^2/(10 \times 0.5 \times 0.5) + (2 - 10 \times 0.5 \times 0.5)^2/(10 \times 0.5 \times 0.5) + (3 - 10 \times 0.5 \times 0.5)^2/(10 \times 0.5 \times 0.5) + (2 - 10 \times 0.5 \times 0.5)^2/(10 \times 0.5 \times 0.5) = 0.4$

Randomisation Example

If we have $N = 5, X = 1, Y = 4$, then we have the sample:

Past tweets					Current tweets				
1	0	0	0	0	1	1	1	1	0

We then randomly divide the 10 tweets into two groups of $N = 5$.

Past tweets					Current tweets				
1	1	0	1	0	0	1	0	1	0

$\rightarrow$

	Topic	No Topic
Past	3	2
Curr	2	3

Giving $\chi^2 = (3 - 10 \times 0.5 \times 0.5)^2/(10 \times 0.5 \times 0.5) + (2 - 10 \times 0.5 \times 0.5)^2/(10 \times 0.5 \times 0.5) + (3 - 10 \times 0.5 \times 0.5)^2/(10 \times 0.5 \times 0.5) + (2 - 10 \times 0.5 \times 0.5)^2/(10 \times 0.5 \times 0.5) = 0.4$

Repeat 1000 times to obtain the χ^2 distribution given H_0.

χ^2 Distribution

Problem

Below is the χ^2 distribution for our previous example. Do we reject H_0 when $\chi^2 = 1.3956$?

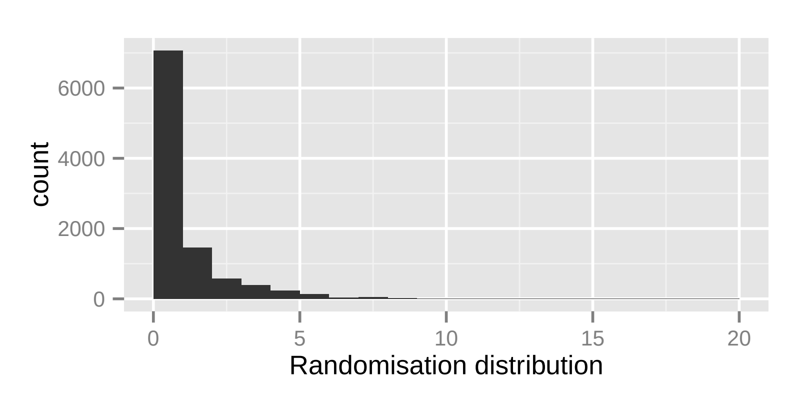


Figure: Distribution of chi squared statistic.

χ^2 test using R

The χ^2 test output in R provides the value of the test statistic and the p value.

```
m0 = matrix(c(181, 584, 201, 564),byrow=TRUE, 2)
chisq.test(m0, simulate.p.value=TRUE)
```

```
##
##  Pearson’s Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  m0
## X-squared = 1.3956, df = NA, p-value = 0.2639
```

Outline

Regression

If we model the count of tweets containing the topic over time, and the model shows us an increase in tweets over time, we have found a trend.

One type of model we can use is a Simple Linear Regression (SLR).

Simple Linear Regression of tweets over time

If Y_t is the count of tweets on a topic in time period t we can examine the model:

$$Y_t = \alpha + \beta t + \varepsilon$$

Where α represents the tweet count at time zero, β is the slope (the change in count per unit time). If β is positive, we have found a trend.
The term ε represents an additive random noise element.

Simple Linear Regression for Twitter data?

For tweets there are problems.

1. SLR assumes that ε is Normally distributed and has constant variance over time. Counts are rarely Normally distributed, and their variance σ^2 is not constant.
2. SLR assumes that all ε are independent. The actual errors might occur in runs i.e. be over the regression line in a group then under the line. This means they are NOT independent.
3. SLR is linear. The trend may not be very linear.

Simple Linear Regression for Twitter data?

For tweets there are problems.

1. SLR assumes that ε is Normally distributed and has constant variance over time. Counts are rarely Normally distributed, and their variance σ^2 is not constant.
2. SLR assumes that all ε are independent. The actual errors might occur in runs i.e. be over the regression line in a group then under the line. This means they are NOT independent.
3. SLR is linear. The trend may not be very linear.

The first can often be fixed with a transformation — typically a log or square root. Taking logs or square roots usually has the effect of making count data more Normally distributed and stabilising the variance.

In fact, for Poisson counts, the square root is the *Variance Stabilising Transformation* and doesn't have an issue with zero counts.

Variance Stabilising Transformation

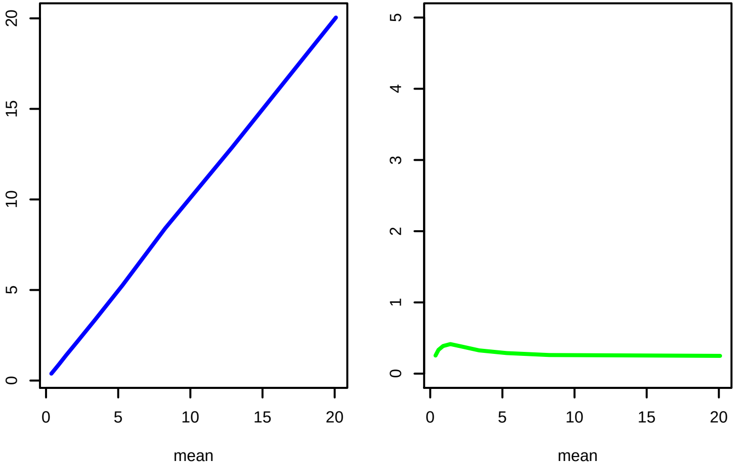


Figure: Variance of 10000 Poisson RV (right=square root) Scales are not equal

Simple Linear Regression for Twitter data?

The second problem is that the counts are probably not independent.

- It is likely that a high tweet count has an impact on the next time period.
- There are ways to deal with this (beyond the scope of this unit)

The third problem is nonlinear trends

- For example, exponential growth in counts?
- To some extent, transformation makes the trend in counts nonlinear
  - Log — exponential
  - Sqrt — quadratic
- This has limited use.
- Other methods are beyond the scope of this unit.

Example of simple linear regression

Data collected by ScraperWiki.

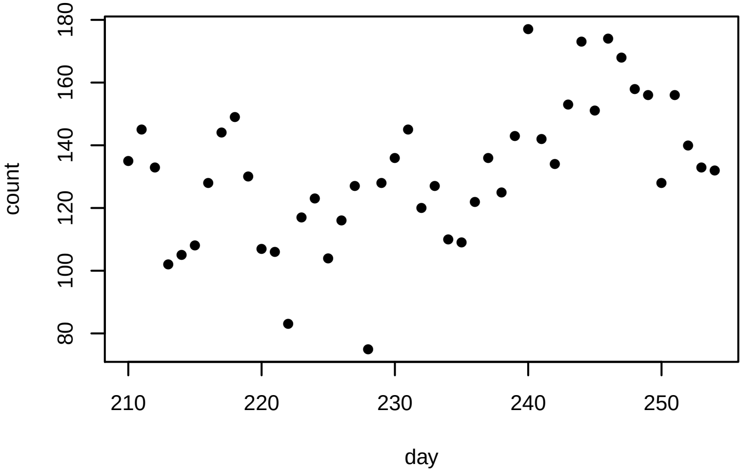


Figure: Counts of #iPhone5 tweets for a period in 2013

Example of simple linear regression

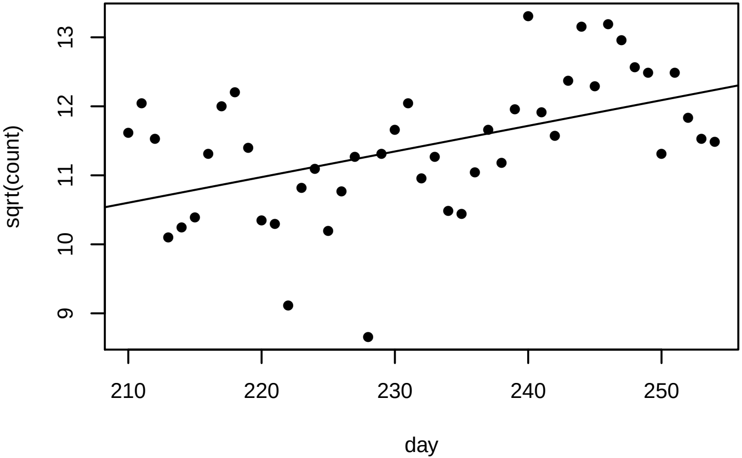


Figure: Square root Counts of #iPhone5 tweets for a period in 2013 and fitted SLR

Example of simple linear regression

Linear regression with R. The *p* value for day shows that there is a trend.

```
m = lm(I(sqrt(count)) ~ day, data = df2)
summary(m)
```

```
##
## Call:
## lm(formula = I(sqrt(count)) ~ day, data = df2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6112 -0.5232  0.0051  0.5408  1.5870
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.80411    2.36352   1.186 0.241973
## day          0.03714    0.01017   3.651 0.000703 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8862 on 43 degrees of freedom
```

Outline

Twitter counts as a time series

If we are counting the number of tweets in a fixed time period, for a sequence of equally spaced times, this is an example of a time series in statistics.

There are many models for time series data that allow for all the issues. (Normality, Independence, nonlinear trend)

Here we will look at a simple (descriptive) technique.

Trend, Seasonal and Irregular components

The basic idea is that a time series can be broken down into 3 components.

$$Y_t = T_t + S_t + \varepsilon_t$$

where

- Y_t is the data
- T_t is the (possibly nonlinear) trend. This represent the movement in the mean over time.
- S_t is the seasonal component. (See below)
- ε_t is the irregular or random component.

Trend, Seasonal and Irregular components

The trend is a smooth component that varies slowly and represents how the mean value of the data changes over time.

The seasonal component consist of any periodic/cyclical behaviour in the data. For example, in temperature data, we expect certain months to be hotter than others in a systematic way.

If the seasonal cycle consists of m time periods, (e.g. $m = 12$ for monthly weather data with an annual seasonal cycle), then we have two properties of the seasonal component.

- $S_{t+m} = S_t$: the seasonal component is periodic
- $\sum_{t=1}^{m} S_t = 0$: the seasonal component over one cycle sums to zero — this means, the mean seasonal effect is zero.

Trend, Seasonal and Irregular components

Other examples of *seasonal* cycles are;

- An annual cycle in monthly recorded data — temperatures, etc. ($m = 12$)
- An annual cycle in quarterly data — often used for sales data. ($m = 4$)
- A weekly cycle in daily recorded data — car traffic, share prices(?), ($m = 7$)
- A daily cycle in hourly data — emergency phone calls, traffic, ... ($m = 24$)

The irregular component represents random variation not explained by trend and seasonal components. It has mean (expected value) zero.

Moving Averages

One way to estimate the trend in the above *model* is to use a moving average.

A moving using a window of $k < n$ points takes an average over the first k points, then moves one space and averages over the k points.

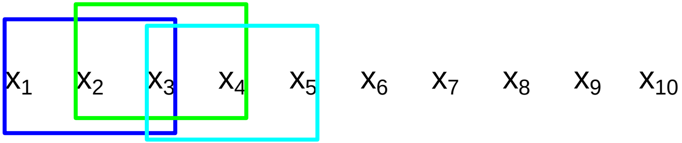


Figure: 3pt moving average

Moving Averages

The central point of the window is then replaced by the moving average.

In mathematical notation this can be written, for an odd sized window, say $k = 2 \times j + 1$

$$Z_t = \frac{1}{k} \sum_{i=-j}^{j} Y_{t+i} \qquad \forall \quad j < t \leq n - j \qquad \text{undefined for other } t$$

For an even sized window $k = 2 \times j$, we do not have an obvious central point so we make it an odd window, but use a half weight at the end points, so the sum of the weights is still k.

$$Z_t = \frac{1}{k} \left(\frac{1}{2} Y_{t-j} + \sum_{i=-j+1}^{j-1} Y_{t+i} + \frac{1}{2} Y_{t+j} \right) \qquad \forall \quad j < t \leq n - j$$

Moving Averages for Trend estimation

Since $\sum_{i=1}^{m} S_i = 0$, if we take a moving average with window size m, we eliminate the seasonal component.

$$Y_t = T_t + S_t + \varepsilon_t$$

So (for $m = 2j + 1$ odd)

$$Z_t = \frac{1}{m}(\sum_{i=-j}^{j} Y_{t+i} = \sum_{i=-j}^{j} T_{t+i} + \sum_{i=-j}^{j} S_{t+i} + \sum_{i=-j}^{j} \varepsilon_{t+i})$$

And the middle term disappears since $S_{t+m} = S_t$ and $\sum_{i=1}^{m} S_i = 0$

Z_t is then taken as an *estimate* of T_t. If T_t is *smooth* it will be a good estimate.

IPhone5 hashtag

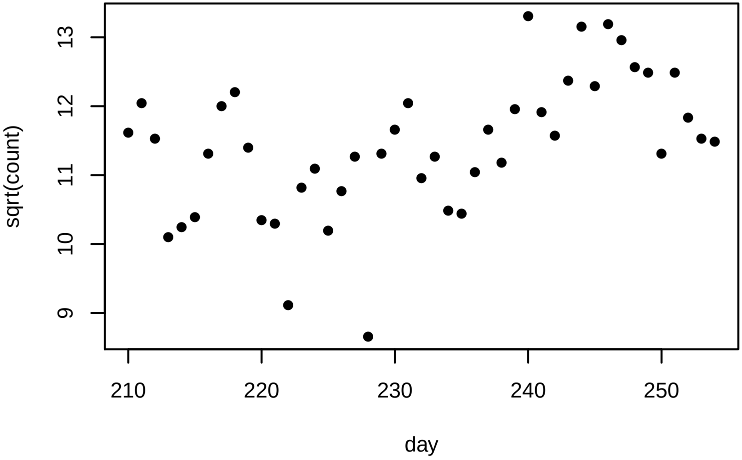
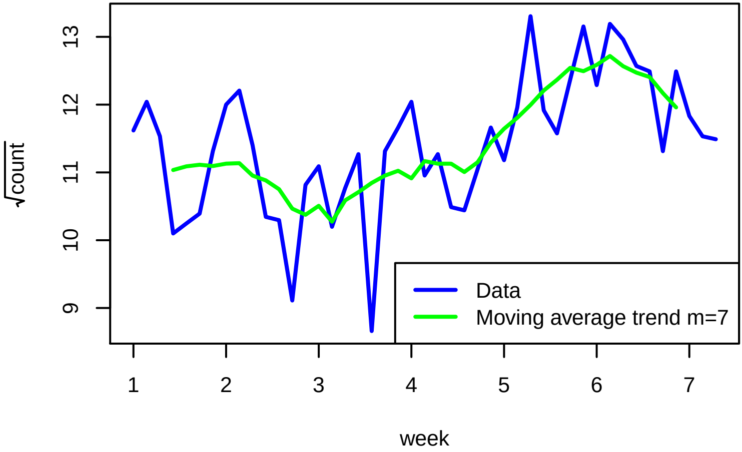


Figure: Square root Counts of #iPhone5 tweets for a period in 2013

IPhone5 hashtag

This data is counts on 45 consecutive days in the middle of 2013.

If there is a seasonal component it is probably weekly, that is, $m = 7$.



IPhone5 hashtag

For this data Y_t the response is the square root of counts of tweets at each day, starting at day 210, finishing at day 254. So for a 7-point moving average we can start at day 213, and go up to day 251.

For example, the trend at day 213, is

$$(Y_{210} + Y_{211} + Y_{212} + Y_{213} + Y_{214} + Y_{215} + Y_{216})/7$$

day	210	211	212	213	214	215	216
count	135	145	133	102	105	108	128

$$T_{213} = (11.62 + 12.04 + 11.53 + 10.10 + 10.25 + 10.39 + 11.31)/7 = 11.03$$

Trend at 214 would be

$$(Y_{211} + Y_{212} + Y_{213} + Y_{214} + Y_{215} + Y_{216} + Y_{217})/7$$

Moving average with an Even window

If m is even, we use the adjusted version, so that the moving average is over an odd number of points

$$T_t = \frac{1}{m}(\frac{1}{2}Y_{t-j} + \sum_{i=-j+1}^{j-1} Y_{t+i} + \frac{1}{2}Y_{t+j})$$

For example, if $m = 4$ in the above data, trend at day 213 becomes

$$(\frac{1}{2}Y_{211} + Y_{212} + Y_{213} + Y_{214} + \frac{1}{2}Y_{215})/4$$

$$T_{213} = (\frac{1}{2} \times 12.04 + 11.53 + 10.10 + 10.25 + \frac{1}{2} \times 10.25)/4 = 10.77$$

So the moving average is actually over 5 points, but the *weights* add up to 4.

Computing moving averages

Example

Compute the trend component in the fifth position of the following sequence, with seasonal cycle of length 6.

7	6	5	4	8	7	5	4	6	6	4	3

Computing moving averages

Example

Compute the trend component in the fifth position of the following sequence, with seasonal cycle of length 6.

7	6	5	4	8	7	5	4	6	6	4	3

Problem

Compute the trend component in the seventh position of the following sequence, with seasonal cycle of length 3.

7	6	5	4	8	7	5	4	6	6	4	3

Seasonal Component

The above has assumed a 7-day seasonal or period component. How do we estimate this component?

Simply, subtract the trend, then average over each time corresponding to the same part of the cycle.

$$\tilde{S}_t = \frac{1}{n_t} \sum_j \left(Y_{t+jm} - T_{t+jm} \right) \qquad \forall \quad t = 1, \ldots, m$$

where the sum is over all j such that $t + jm$ is in the data, and n_t is the number of terms that occur.

Unfortunately, these $\tilde{S}_t$ do not necessary sum to zero, which a seasonal component should do, so we subtract the mean, and set

$$S_t = \tilde{S}_t - \frac{1}{m} \sum_{j=1}^{m} \tilde{S}_j \qquad \forall \quad t = 1, \ldots, m$$

Seasonal Component

One way to think about this is to configure the data as a table/matrix, with m rows and n/m columns, with *time* running down each column.

The seasonal is the average along the rows after the trend is subtracted.

Iphone5 hashtag

The data has $n = 45$ data points and $m = 7$ for a weekly cycle.

1	135	144	123	145	125	151	140
2	145	149	104	120	143	174	133
3	133	130	116	127	177	168	132
4	102	107	127	110	142	158	
5	105	106	75	109	134	156	
6	108	83	128	122	153	128	
7	128	117	136	136	173	156	

Iphone5 hashtag

We are using the square roots of the counts.

Data

1	11.62	12.00	11.09	12.04	11.18	12.29	11.83
2	12.04	12.21	10.20	10.95	11.96	13.19	11.53
3	11.53	11.40	10.77	11.27	13.30	12.96	11.49
4	10.10	10.34	11.27	10.49	11.92	12.57	
5	10.25	10.30	8.66	10.44	11.58	12.49	
6	10.39	9.11	11.31	11.05	12.37	11.31	
7	11.31	10.82	11.66	11.66	13.15	12.49	

Iphone5 hashtag

Trend is a 7 point moving average

Trend

1		11.13	10.51	10.91	11.64	12.59
2		11.14	10.27	11.17	11.81	12.72
3		10.95	10.59	11.13	12.00	12.57
4	11.04	10.88	10.71	11.13	12.21	12.47
5	11.09	10.75	10.85	11.01	12.37	12.41
6	11.11	10.47	10.95	11.15	12.54	12.17
7	11.09	10.38	11.02	11.44	12.49	11.96

Iphone5 hashtag

WESTERN SYDNEY UNIVERSITY

Difference

1		0.87	0.58	1.13	-0.46	-0.30
2		1.07	-0.08	-0.21	0.15	0.47
3		0.45	0.18	0.14	1.31	0.39
4	-0.94	-0.54	0.56	-0.64	-0.29	0.10
5	-0.84	-0.46	-2.18	-0.57	-0.79	0.08
6	-0.72	-1.35	0.36	-0.10	-0.17	-0.86
7	0.22	0.44	0.64	0.22	0.66	0.53

Iphone5 hashtag

WESTERN SYDNEY UNIVERSITY

$\tilde{S}_t$

1	2	3	4	5	6	7
0.364	0.282	0.495	-0.291	-0.793	-0.475	0.452

$S_t = \tilde{S}_t - \frac{1}{m} \sum_{j=1}^{m} \tilde{S}_j$

1	2	3	4	5	6	7
0.359	0.277	0.490	-0.296	-0.798	-0.479	0.447

Irregular component

The irregular component is simply everything that is left over.

$$Y_t - T_t - S_t$$

Iphone5 hashtag

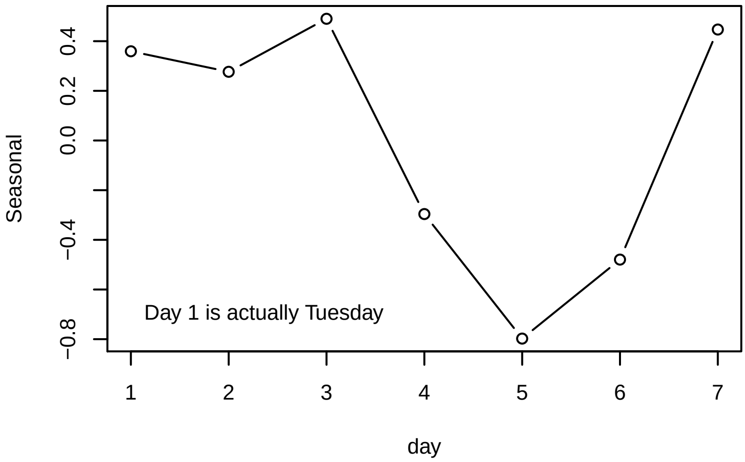


Figure: Seasonal component for iphone5 hashtag

Seasonal problem

Problem

Given the following data with seasonal cycle of length 4:

7	6	5	4	8	7	5	4	6	6	4	3

with seasonal components:

1.390625	1.015625	NA	-1.671875

Compute the third seasonal component.

Summary

So in this lecture we looked at

- Twitter's method for finding *Twitter Trends*
- Revised *Simple Linear Regression* as a simple way to look at trends
- Saw how moving averages can be used to estimate trends and seasonality

Next time, *Events*