Time Series

To infinity and beyond….
by Ralf Martin (r.martin@imperial.ac.uk)
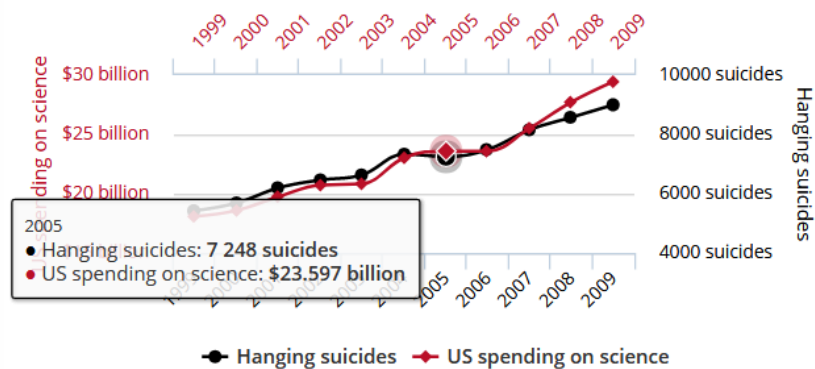
Objectives of this lecture

- Time Series data: Different data points represent different points in time

- This introduces some additional challenges

- We will discuss how to deal with those

What's the challenge of time series data?

US spending on science, space, and technology correlates with Suicides by hanging, strangulation and suffocation — Correlation: 99.79% (r=0.99789126)
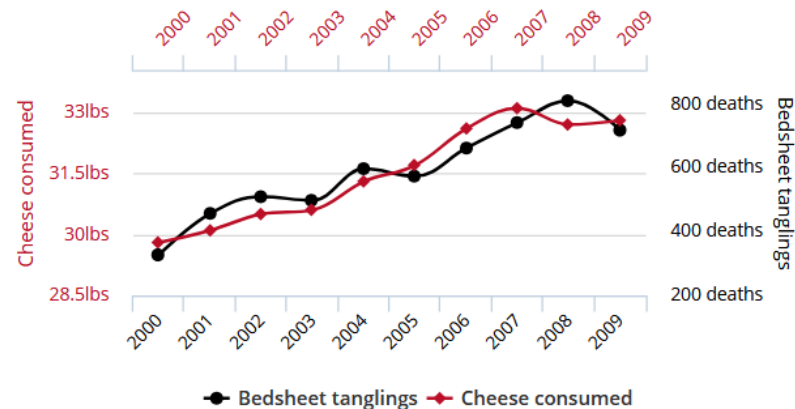
2005 • Hanging suicides: 7 248 suicides • US spending on science: $23.597 billion

Data sources: U.S. Office of Management and Budget and Centers for Disease Control & Prevention

Per capita cheese consumption correlates with Number of people who died by becoming tangled in their bedsheets — Correlation: 94.71% (r=0.947091)
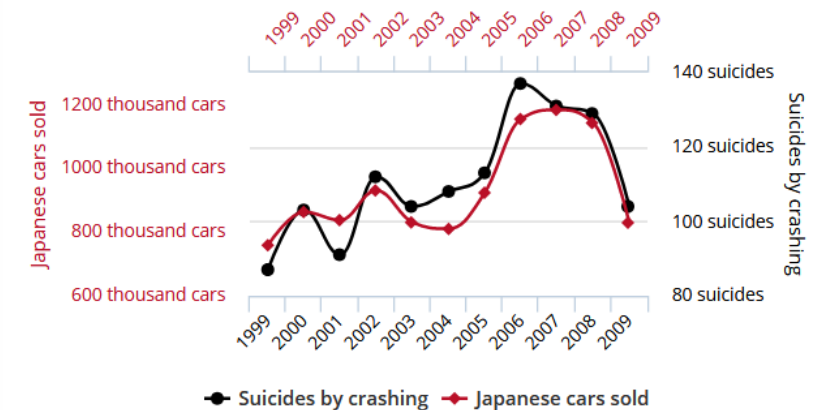
Data sources: U.S. Department of Agriculture and Centers for Disease Control & Prevention

Japanese passenger cars sold in the US correlates with Suicides by crashing of motor vehicle — Correlation: 93.57% (r=0.935701)
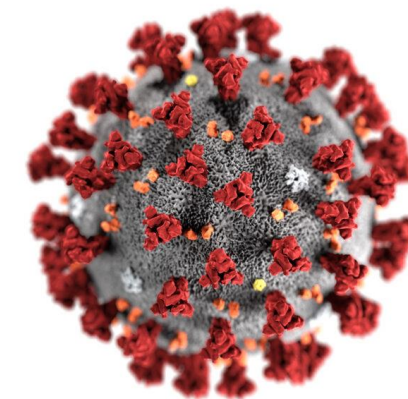
Data sources: U.S. Bureau of Transportation Statistics and Centers for Disease Control & Prevention

Can you spot the problem?

Time becomes a confounding variable
Non-stationary: characteristics of data vary with time

COVID vs GDP

```
head(df)
```

```
##         week  WEI   Index cases deaths      lnindex lockshare
## 1 2008-01-05 1.42 1.00000     0      0 0.0000000000         0
## 2 2008-01-12 1.46 1.00028     0      0 0.0002799608         0
## 3 2008-01-19 1.40 1.00055     0      0 0.0005498488         0
## 4 2008-01-26 0.96 1.00073     0      0 0.0007297337         0
## 5 2008-02-02 0.73 1.00088     0      0 0.0008796130         0
## 6 2008-02-09 0.78 1.00103     0      0 0.0010294699         0
```

What you think is going to happen?

In 100K of cases

More COVID = more GDP?
100K more = 5% more GDP?
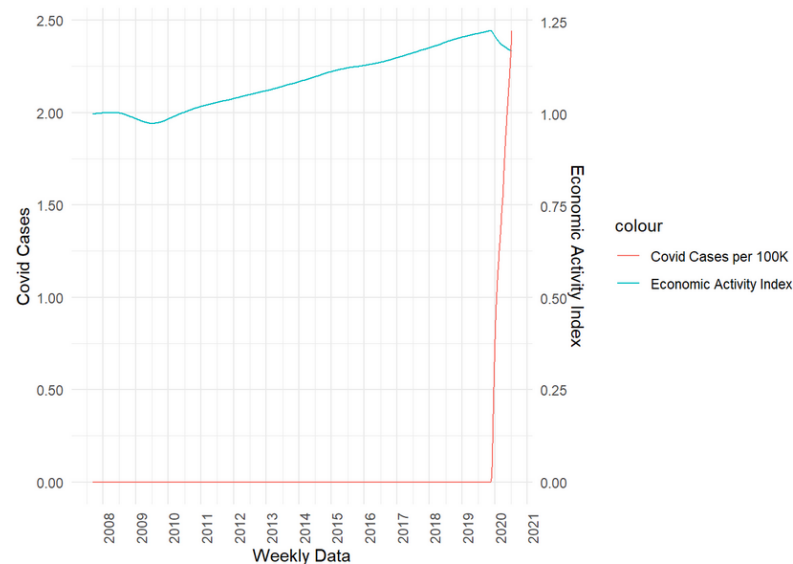

```
lm(lnindex~cases,df) %>% summary()
```

```
##
## Call:
## lm(formula = lnindex ~ cases, data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.108375 -0.064942 -0.002043   0.055      0.121388
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.082359   0.002731  30.156  < 2e-16 ***
## cases       0.050576   0.007800   6.484 1.74e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06928 on 669 degrees of freedom
## Multiple R-squared:  0.05913,    Adjusted R-squared:  0.05772
## F-statistic: 42.04 on 1 and 669 DF,  p-value: 1.736e-10
```

Taking control of time….with a timeline

```
df=df %>% mutate(t=1:n())
lm(lnindex~cases+t,df) %>% summary()
```

```
##
## Call:
## lm(formula = lnindex ~ cases + t, data = df)
##
## Residuals:
##       Min       1Q    Median       3Q      Max
## -0.024859 -0.004965 -0.001175  0.003861  0.038124
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.850e-02  9.170e-04  -41.98   <2e-16 ***
## cases       -2.262e-02  1.393e-03  -16.23   <2e-16 ***
## t            3.752e-04  2.466e-06  152.11   <2e-16 ***
## ---
## Signif. c     s:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual        or: 0.01161 on 668 degrees of freedom
## Multiple         6, Adjusted R-squared:  0.9735
## F-statist               668 DF,  p-value: < 2.2e-16
```

100k more cases =
2.26% lower GDP

What if time is not linear?

- Seasonal effects
- Recessions
- Natural disasters

- Political turmoil
- War
- Pandemic

Panel data to the rescue

```
head(statsbyweek %>% arrange(state,week))
```

```
## # A tibble: 6 x 9
## # Groups:   state [1]
##   state    week       hoax tweets cases deaths hoaxsh Dcases Ddeaths
##   <chr>    <date>     <int>  <int> <int>  <int>  <dbl>  <int>   <int>
## 1 Alabama  2020-03-15     4   1503    51      0  0.266     NA      NA
## 2 Alabama  2020-03-22    62   4198   386      1  1.48     335       1
## 3 Alabama  2020-03-29    14   5218  1108     28  0.268    722      27
## 4 Alabama  2020-04-05    12   4793  2498     67  0.250   1390      39
## 5 Alabama  2020-04-12     9   4486  4241    123  0.201   1743      56
## 6 Alabama  2020-04-19     6   3570  5610    201  0.168   1369      78
```
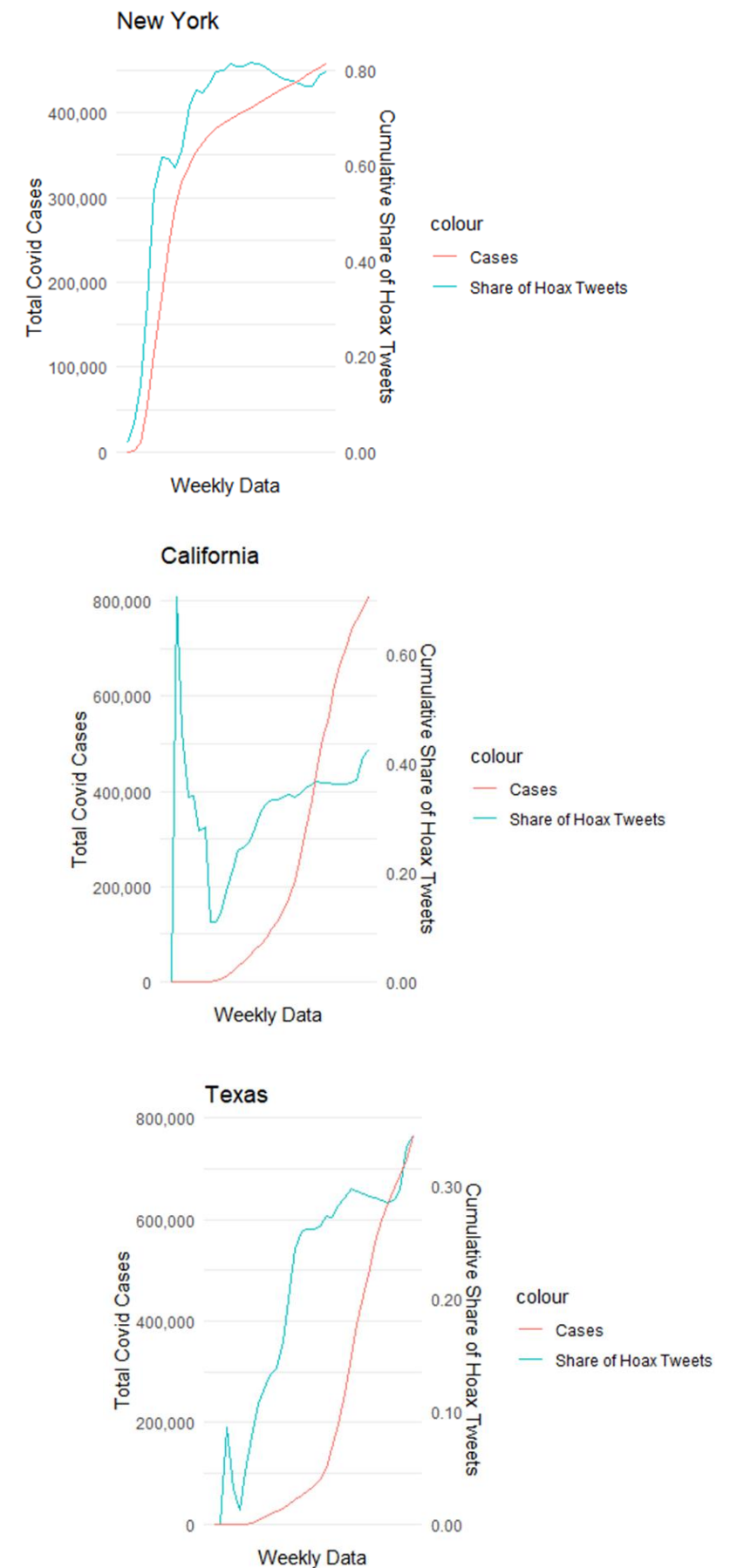
```
statsbyweek %>% group_by(state) %>% summarise(n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 50 x 2
##    state        `n()`
##    <chr>        <int>
##  1 Alabama         29
##  2 Alaska          29
##  3 Arizona         36
##  4 Arkansas        30
##  5 California      36
##  6 Colorado        30
##  7 Connecticut     30
##  8 Delaware        30
##  9 Florida         31
## 10 Georgia         31
## # ... with 40 more rows
```

Multiple periods for the same cross section unit

New York

California

Texas

Panel data

	state	week	hoax	tweets	cases	deaths	hoaxsh	Dcases	Ddeaths	cumhoax	cumtweets	cumhoaxsh	Dcumhoaxsh
1	Alabama	2020-03-08	1	443	12	0	0.22573363	*NA*	*NA*	1	443	0.22573363	*NA*
2	Alabama	2020-03-15	13	2098	131	0	0.61963775	119	0	14	2541	0.55096419	3.252306e-01
3	Alabama	2020-03-22	57	5824	720	4	0.97870879	589	4	71	8365	0.84877466	2.978105e-01
4	Alabama	2020-03-29	10	4750	1632	44	0.21052632	912	40	81	13115	0.61761342	-2.311612e-01
5	Alabama	2020-04-05	16	4477	3262	93	0.35738218	1630	49	97	17592	0.55138699	-6.622643e-02
6	Alabama	2020-04-12	6	4180	4723	151	0.14354067	1461	58	103	21772	0.47308470	-7.830230e-02
7	Alabama	2020-04-19	7	3294	6213	213	0.21250759	1490	62	110	25066	0.43884146	-3.424324e-02
8	Alabama	2020-04-26	5	2435	7611	289	0.20533881	1398	76	115	27501	0.41816661	-2.067485e-02
9	Alabama	2020-05-03	20	593	9668	390	3.37268128	2057	101	135	28094	0.48052965	6.236304e-02
10	Alabama	2020-05-10	21	429	11674	485	4.89510490	2006	95	156	28523	0.54692704	6.639739e-02
11	Alabama	2020-05-17	5	816	14149	549	0.61274510	2475	64	161	29339	0.54875763	1.830585e-03
27	Alabama	2020-09-06	4	327	137646	2330	1.22324159	5843	75	205	39056	0.52488754	5.896404e-03
28	Alabama	2020-09-13	21	381	144164	2437	5.51181102	6518	87	226	39437	0.57306590	4.817856e-02
29	Alabama	2020-09-20	0	280	151591	2506	0.00000000	7427	69	226	39717	0.56902586	-4.040045e-03
30	Alaska	2020-03-08	1	59	1	0	1.69491525	*NA*	*NA*	1	59	1.69491525	*NA*
31	Alaska	2020-03-15	5	300	21	0	1.66666667	20	0	6	359	1.67130919	-2.360606e-02
32	Alaska	2020-03-22	1	776	102	1	0.12886598	81	1	7	1135	0.61674009	-1.054569e+00
33	Alaska	2020-03-29	38	863	169	3	4.40324450	67	2	45	1998	2.25225225	1.635512e+00
34	Alaska	2020-04-05	7	993	255	6	0.70493454	86	3	52	2991	1.73854898	-5.137033e-01
35	Alaska	2020-04-12	7	805	312	7	0.86956522	57	1	59	3796	1.55426765	-1.842813e-01
36	Alaska	2020-04-19	6	842	337	7	0.71258907	25	0	65	4638	1.40146615	-1.528015e-01
37	Alaska	2020-04-26	3	741	363	7	0.40485830	26	0	68	5379	1.26417550	-1.372907e-01
38	Alaska	2020-05-03	1	105	377	8	0.95238095	14	1	69	5484	1.25820569	-5.969808e-03
39	Alaska	2020-05-10	3	109	392	8	2.75229358	15	0	72	5593	1.28732344	2.911775e-02

Panel data example

```
lm(cases~cumhoaxsh, statsbyweek) %>% summary()

##
## Call:
## lm(formula = cases ~ cumhoaxsh, data = statsbyweek)
##
## Residuals:
##     Min      1Q  Median      3Q      M
## -294697  -38574  -24113    3608   7   32
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    22072       4127   5.348 1.02e-07 ***
## cumhoaxsh     121120      11220  10.796  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 103100 on 1518 degrees of freedom
## Multiple R-squared:  0.0713, Adjusted R-squared:  0.07069
## F-statistic: 116.5 on 1 and 1518 DF,  p-value: < 2.2e-16
```

Hoax share up by 1 percentage point means 121120 more cases

```
lm(cases~cumhoaxsh+factor(week), statsbyweek) %>% summary()

##
## Call:
## lm(formula = cases ~ cumhoaxsh + factor(week), data =
## statsbyweek)
##
## Residuals:
##     Min      1Q  Median      3Q      Ma
## -182000  -36269   -9070    5947  6656
##
## Coefficients:
##                       Estimate Std   rror t value Pr(>|t|)
## (Intercept)                  1     55848   0.000   1.0000
## cumhoaxsh                72056     11578   6.224 6.31e-10 ***
## factor(week)2020-01-26  -10172     70662  -0.144   0.8856
## factor(week)2020-02-02   -5521     68406  -0.081   0.9357
## factor(week)2020-02-09   -3486     66754  -0.052   0.9584
## factor(week)2020-02-16   -3073     65490  -0.047   0.9626
## factor(week)2020-02-23  -17368     63738  -0.272   0.7853
## factor(week)2020-03-01   -3010     58410  -0.052   0.9589
## factor(week)2020-03-08   -4210     57537  -0.073   0.9417
## factor(week)2020-03-15   -5968     57509  -0.104   0.9174
## factor(week)2020-03-22   -4942     57512  -0.086   0.9315
```

Smaller effect when controlling for time (week) effects

```
lm(cases~cumhoaxsh+factor(state)+factor(week),statsbyweek) %>% summary()

##
## Call:
## lm(formula = cases ~ cumhoaxsh + factor(state) + factor(week),
##     data = statsbyweek)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -244737  -21562    1446   21041  463246
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -109303      41227  -2.651 0.008107 **
## cumhoaxsh                 61619      19128   3.221 0.001304 **
## factor(state)Alaska      -96560      22677  -4.258 2.20e-05 ***
## factor(state)Arizona      34689      16951   2.046 0.040899 *
## factor(state)Arkansas     -9094      18337  -0.496 0.620015
## factor(state)California  224047      17120  13.087  < 2e-16 ***
## factor(state)Colorado     -8425      17692  -0.476 0.634008
## factor(state)Connecticut   4626      18572   0.249 0.803343
## factor(state)Delaware    -31191      17875  -1.745 0.081207 .
## factor(state)Florida     206617      17752  11.639  < 2e-16 ***
```
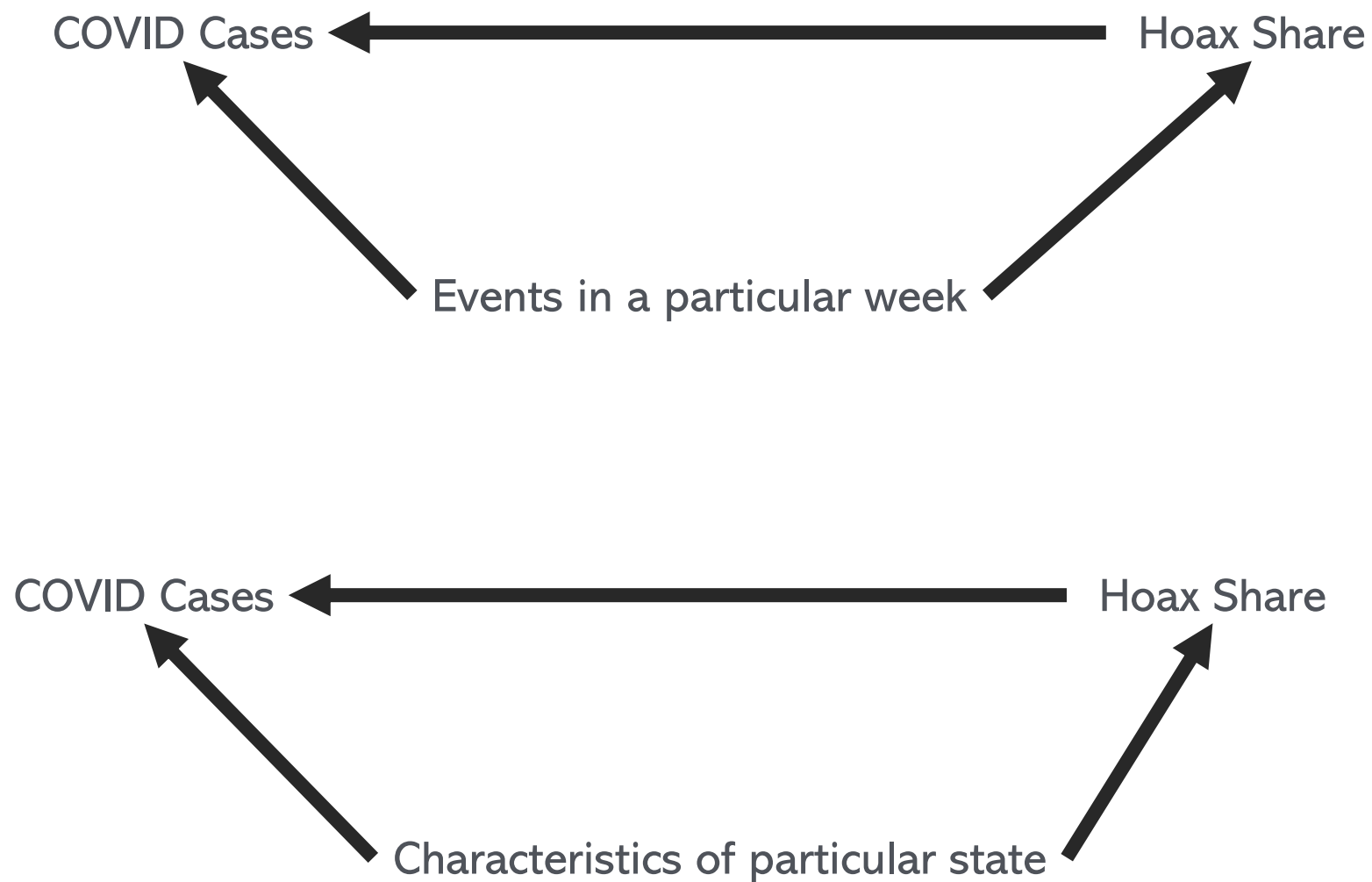
Also controlling for state

e.g.: more rural and less densely populated states have also less people engaged in hoax conspiracies

Example issue: suppose in some weeks there are school holidays (and hence a lower number of covid cases). Also suppose that hoax tweeters are more active over the holidays

Time and cross sectional unit as co-founder

We include dummy for every state and every year
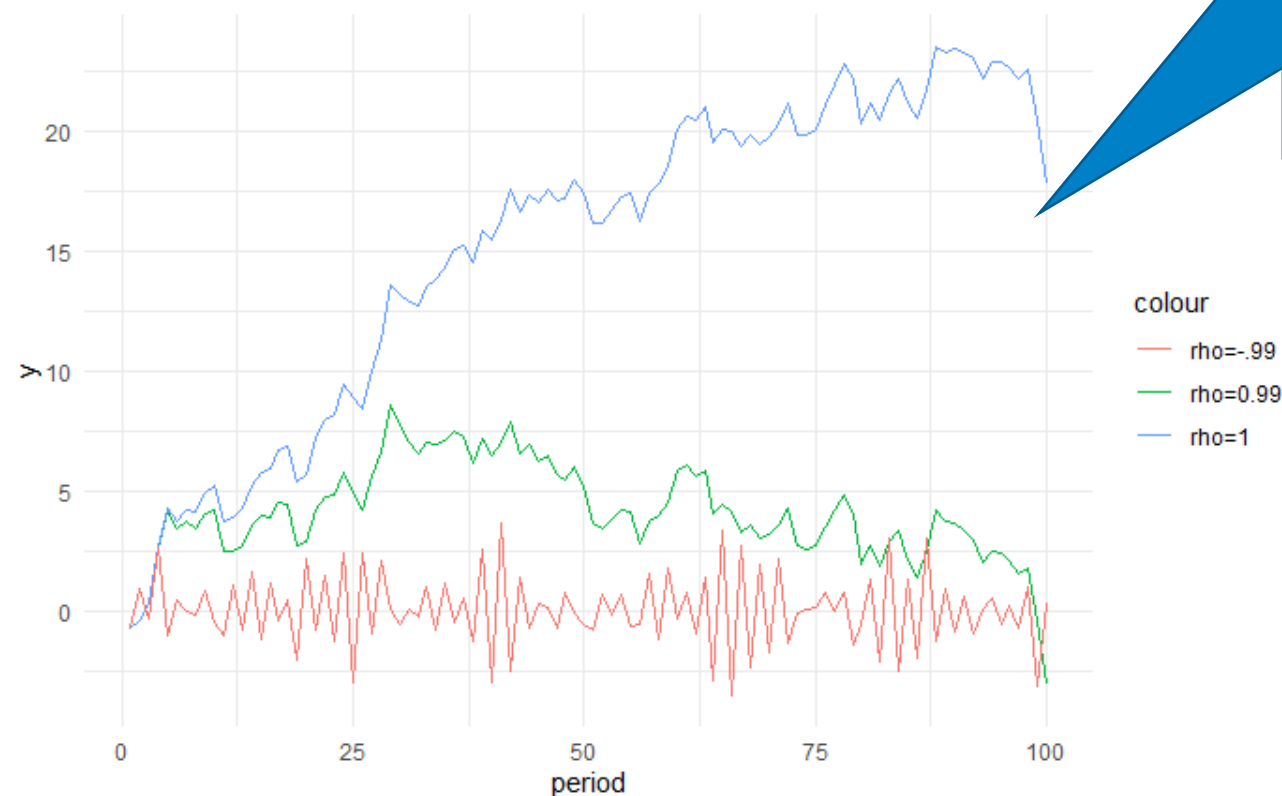Also known as **fixed effects**

Autoregression

- A particular concern in time series is the possibility that observations are correlated over time

- Simplest way to model this is via an Auto regression:

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

Y_{t-1} becomes the X variable
We can do normal OLS as long as $-1 < \rho < 1$

- With $\rho = 1$ we have non-stationarity because of path dependence
- The series can wander off into any direction and neve come back

- If that happens OLS is no longer un-biased (different observations are too related to each other)
- Also: if you are interested in $Y = \beta X$ (i.e. the casual effect of X on Y) you get biased estimates

- Random Walk
- Of course we don't know if this is the case in our data before we start any analysis

Need to test $\hat{\rho} = 1$ but cannot use simple t or F test because under H0 it would be biased

Dickey-Fuller test to the rescue

Rewrite original model by subtracting Y_{t-1} on both sides of the model equation:

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

$$\Downarrow$$

$$Y_t - Y_{t-1} = \Delta Y_t = \beta_0 + \underbrace{(\rho - 1)}_{=\delta} Y_{t-1} + \epsilon_t$$

Testing for a random walk (aka unit root) now boils down to
H0: δ=0
H1: δ<0 i.e. stationary process

- We cannot just compare the implied test statistic to a normal t-table
- Luckily R will help us

R to the rrrrrescue

```
library(urca)
 ur.df(df$lnindex) %>% summary()

##
## ###############################################
## # Augmented Dickey-Fuller Test Unit Root Test #
## ###############################################
##
## Test regression none
##
##
## Call:
## lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -8.343e-04 -2.643e-05  4.240e-06  4.131e-05  1.922e-04
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## z.lag.1    -2.155e-05  2.703e-05  -0.797    0.426
## z.diff.lag  9.924e-01  5.596e-03 177.334   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1
##
## Residual standard error: 7.414e-05 on 667 degrees of freedom
## Multiple R-squared:  0.9812, Adjusted R-squared:  0.9811
## F-statistic: 1.739e+04 on 2 and 667 DF,  p-value: < 2.2e-16
##
##
## Value of test-statistic is: -0.7971
##
## Critical values for test statistics:
##      1pct  5pct 10pct
## tau1 -2.58 -1.95 -1.62
```

We cannot reject unit root because -0.7971>-1.95

Getting rid of unit roots – Economic Activity index

```
ur.df(diff(df$lnindex,1),type="none",lags=1) %>% summary()
```

```
##
## #############################################
## # Augmented Dickey-Fuller Test Unit Root Test #
## #############################################
##
## Test regression none
##
##
## Call:
## lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)
##
## Residuals:
##         Min         1Q      Median         3Q        Max
## -8.236e-04 -3.079e-05   3.980e-06  4.133e-05  1.963e-04
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## z.lag.1     -0.010977   0.005233  -2.098   0.0363 *
## z.diff.lag   0.195464   0.038082   5.133 3.75e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.28e-05 on 666 degrees of freedom
## Multiple R-squared:  0.04215,    Adjusted R-squared:  0.03927
## F-statistic: 14.65 on 2 and 666 DF,  p-value: 5.92e-07
##
##
## Value of test-statistic is: -2.0976
##
## Critical values for test statistics:
##        1pct   5pct  10pct
## tau1  -2.58  -1.95  -1.62
```

- Differencing: $\Delta y_t = y_t - y_{t-1}$

- Checking that differenced series is not unit rood

We can reject unit root (at least at 5%)

Revisiting COVID vs GDP

```
df=df %>% arrange(week) %>% mutate( Dlnindex=lnindex - dplyr::lag(lnindex) ,Llnindex=dplyr::lag(lnindex) )


lm(Dlnindex~cases+t,df) %>% summary()

##
## Call:
## lm(formula = Dlnindex ~ cases + t, data = df)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.175e-03 -6.473e-05  7.992e-05  1.627e-04  1.327e-03
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.729e-04  3.019e-05   5.726 1.55e-08 ***
## cases       -9.290e-04  4.573e-05 -20.314  < 2e-16 ***
## t            3.831e-07  8.113e-08   4.723 2.84e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0003811 on 667 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3856, Adjusted R-squared:  0.3838
## F-statistic: 209.3 on 2 and 667 DF,  p-value: < 2.2e-16
```

100k more cases = 0. 0929% lower GDP…

Lagging

```
df=df %>% arrange(week) %>% mutate( Dlnindex=lnindex - dplyr::lag(lnindex) ,Llnindex=dplyr::lag(lnindex) )


lm(Dlnindex~cases+t,df) %>% summary()

##
## Call:
## lm(formula = Dlni
##
## Residuals:
##        Min
## -2.175e-03 -6.473
##
## Coefficients:
##              Est
## (Intercept)  1.72
## cases       -9.29
## t            3.83
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual
##  (1 obs
## Multiple
## F-stati
```

Note how to create lagged values: use the dplyr::lag() function

	week	WEI	Index	cases	deaths	lnindex	lockshare	t	Dlnindex	Llnindex
1	2008-01-05	1.42	1.00000	0	0	0.0000000000	0	1	NA	NA
2	2008-01-12	1.46	1.00028	0	0	0.0002799608	0	2	2.799608e-04	0.0000000000
3	2008-01-19	1.40	1.00055	0	0	0.0005498488	0	3	2.698880e-04	0.0002799608
4	2008-01-26	0.96	1.00073	0	0	0.0007297337	0	4	1.798849e-04	0.0005498488
5	2008-02-02	0.73	1.00088	0	0	0.0008796130	0	5	1.498793e-04	0.0007297337
6	2008-02-09	0.78	1.00103	0	0	0.0010294699	0	6	1.498569e-04	0.0008796130
7	2008-02-16	1.18	1.00125	0	0	0.0012492194	0	7	2.197495e-04	0.0010294699
8	2008-02-23	0.91	1.00143	0	0	0.0014289785	0	8	1.797591e-04	0.0012492194
9	2008-03-01	0.95	1.00161	0	0	0.0016087053	0	9	1.797268e-04	0.0014289785
10	2008-03-08	0.52	1.00171	0	0	0.0017085396	0	10	9.983428e-05	0.0016087053

Lagging with panel data

```
statsbyweek = statsbyweek %>%

        arrange(state,week) %>%

        group_by(state) %>%

        mutate(Dcases=cases-dplyr::lag(cases,1),
               Ddeaths=deaths-dplyr::lag(deaths,1))
```


	state	week	hoax	tweets	cases	deaths	hoaxsh	Dcases	Ddeaths
1	Alabama	2020-03-08	1	443	12	0	0.22573363	NA	NA
2	Alabama	2020-03-15	13	2098	131	0	0.61963775	119	0
3	Alabama	2020-03-22	57	5824	720	4	0.97870879	589	4
4	Alabama	2020-03-29	10	4750	1632	44	0.21052632	912	40
5	Alabama	2020-04-05	16	4477	3262	93	0.35738218	1630	49
6	Alabama	2020-04-12	6	4180	4723	151	0.14354067	1461	58
7	Alabama	2020-04-19	7	3294	6213	213	0.21250759	1490	62
8	Alabama	2020-04-26	5	2435	7611	289	0.20533881	1398	76
9	Alabama	2020-05-03	20	593	9668	390	3.37268128	2057	101
10	Alabama	2020-05-10	21	429	11674	485	4.89510490	2006	95
11	Alabama	2020-05-17	5	816	14149	549	0.61274510	2475	64
12	Alabama	2020-05-24	3	515	17359	618	0.58252427	3210	69
29	Alabama	2020-09-20	0	280	151591	2506	0.00000000	7427	69
30	Alaska	2020-03-08	1	59	1	0	1.69491525	NA	NA
31	Alaska	2020-03-15	5	300	21	0	1.66666667	20	0
32	Alaska	2020-03-22	1	776	102	1	0.12886598	81	1
33	Alaska	2020-03-29	38	863	169	3	4.40324450	67	2
34	Alaska	2020-04-05	7	993	255	6	0.70493454	86	3
35	Alaska	2020-04-12	7	805	312	7	0.86956522	57	1
36	Alaska	2020-04-19	6	842	337	7	0.71258907	25	0
37	Alaska	2020-04-26	3	741	363	7	0.40485830	26	0
38	Alaska	2020-05-03	1	105	377	8	0.95238095	14	1
39	Alaska	2020-05-10	3	109	392	8	2.75229358	15	0

Summary

- Time series can be easy
- But you need to worry about how stationary your series is
- If the series clearly grows or shrinks continuously definitely include a time trend
- However, even if it doesn't grow (or shrink) the series might contain a unit root
- If that's the case a time trend is not enough
- Use the Dickey Fuller Test to make sure you are dealing with a stationary series
- If not take first difference and check Dickey Fuller again
- Reliable causal estimators you can only get from a stationary series.
- If you have panel data you can control for specific time effects (and cross sectional unit effects)

Extra
Slides

Imperial College
Business School

Imperial means
Intelligent Business

Other considerations

```
lm(Dlnindex~cases+lockshare+t,df) %>% summary()

##
## Call:
## lm(formula = Dlnindex ~ cases + lockshare + t, data = df)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -9.141e-04 -1.092e-04  2.947e-05  1.456e-04  1.104e-03
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.209e-04  2.266e-05    5.335 1.31e-07 ***
## cases         5.646e-04  7.334e-05    7.698 5.00e-14 ***
## lockshare    -2.772e-05  1.205e-06  -23.014  < 2e-16 ***
## t             6.278e-07  6.152e-08   10.205  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0002847 on 666 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.6578, Adjusted R-squared:  0.6562
## F-statistic: 426.7 on 3 and 666 DF,  p-value: < 2.2e-16
```

- lockshare=percentage of population in lockdown
- If 100% of US population go into lockdown GDP goes down by -0.2772% (seems low..more research needed)

More lags AR(2)?

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + u_t.$$

Stationarity now requires

$$\beta_1 + \beta_2 < 1$$

while

$$\beta_1 + \beta_2 = 1$$

$$Y_t - Y_{t-1} = \beta_0 + (\beta_1 + \beta_2 - 1)Y_{t-1} - \beta_2(Y_{t-1} - Y_{t-2}) + \epsilon_t$$

We can test this again using the coefficient on Y_{t-1}

More lags and trend?

$$Y_t - Y_{t-1} = \beta_0 + (\beta_1 + \beta_2 - 1)Y_{t-1} - \beta_2(Y_{t-1} - Y_{t-2}) + \rho t + \epsilon_t$$

Further reading

- On time fixed effects: Hanck et al Chapter 10.4
- Unit roots: Hanck et al Chapter 14.7