Time Series

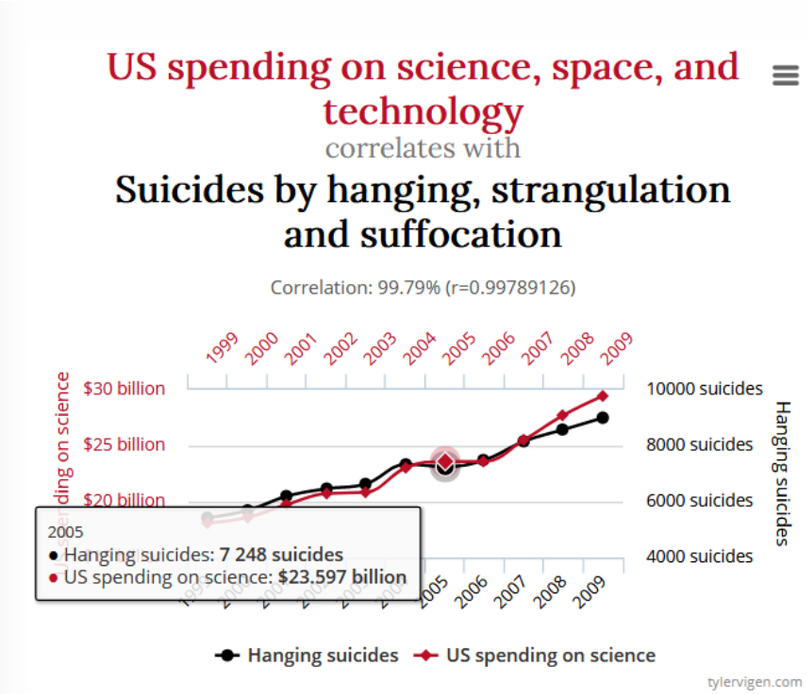
To infinity and beyond….

by Ralf Martin (r.martin@imperial.ac.uk)

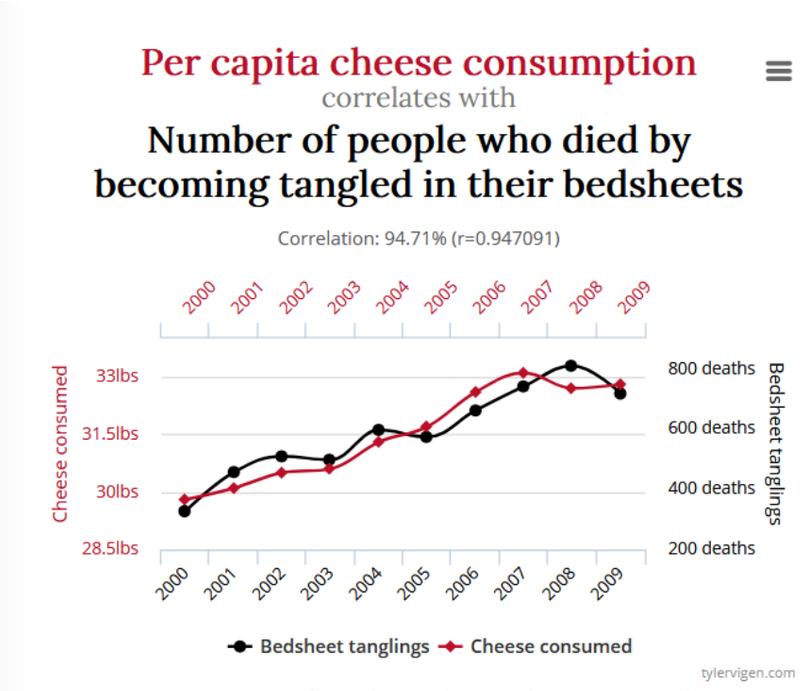
Objectives of this lecture

- Time Series data: Different data points represent different points in time
- This introduces some additional challenges
- We will discuss how to deal with those

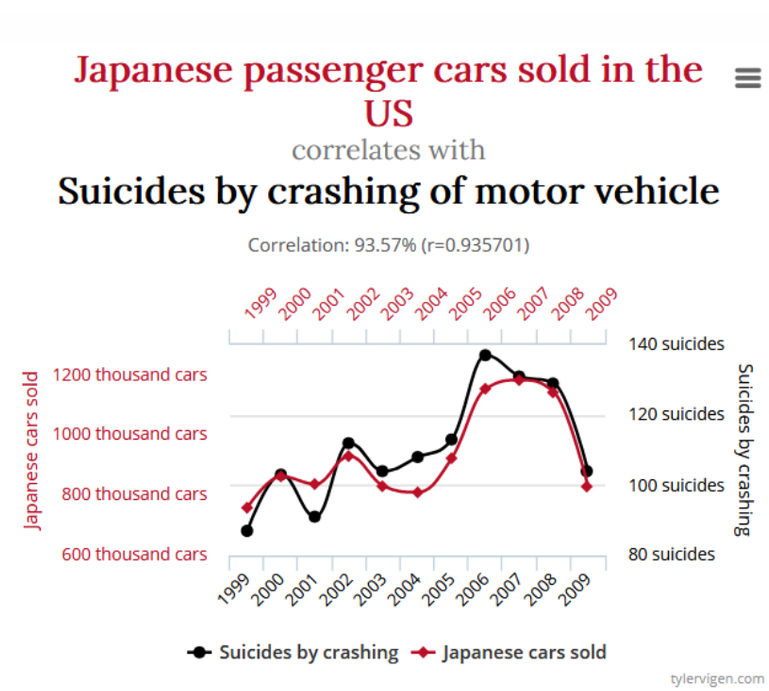
What's the challenge of time series data?



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



Data sources: U.S. Department of Agriculture and Centers for Disease Control & Prevention

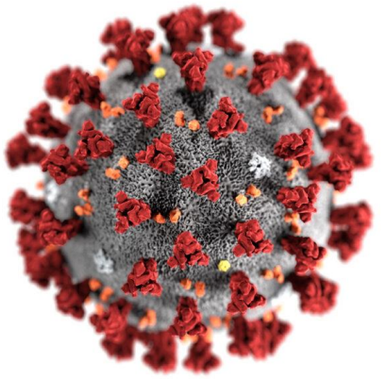


Data sources: U.S. Bureau of Transportation Statistics and Centers for Disease Control & Prevention

Can you spot the problem?

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

COVID vs GDP

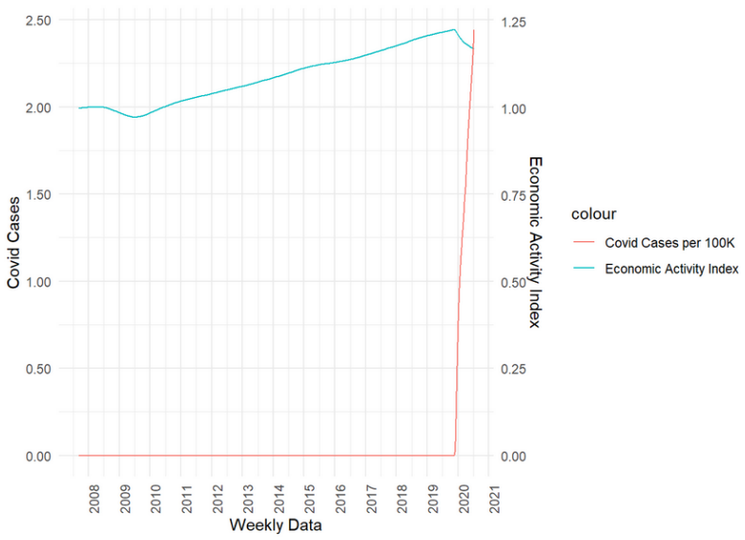


What you think is going to happen?

```
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
```

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[illegible]1  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        error: 0.01161 on 668 degrees of freedom
## Multiple        .9736, Adjusted R-squared:  0.9735
## F-statist          2 and 668 DF,  p-value: < 2.2e-16
```

100k more cases = 2.2% 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))
```

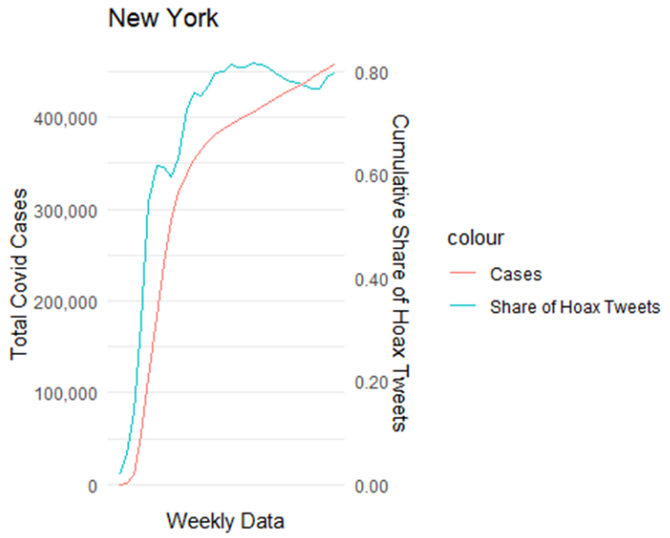
```
## # A tibble: 6 x 9
## # Groups:   state [1]
##   state   week        hoax tweets cases deaths hoaxsh Dcases Ddeaths
##   <chr>   <date>     <int>  <int> <int>  <int>  <dbl>  <int>   <int>
## 1 Alabama 2020-03-15     4   1503    51      0  0.266     NA      NA
## 2 Alabama 2020-03-22    62   4198   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
```

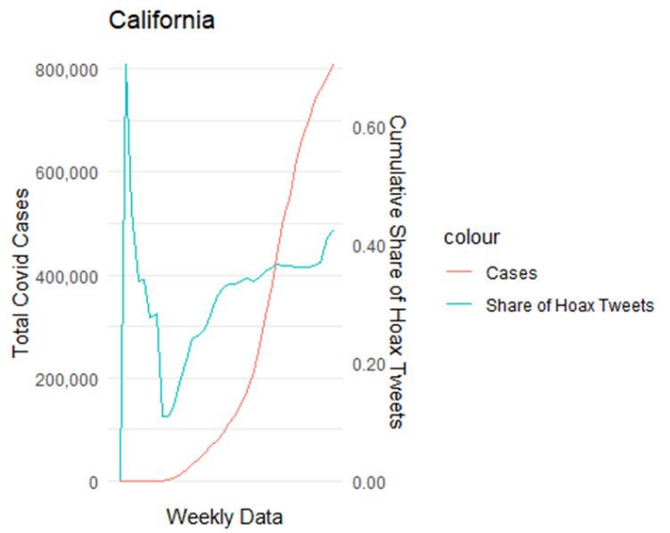
```
statsbyweek %>% group_by(state) %>% summarise(n())
```

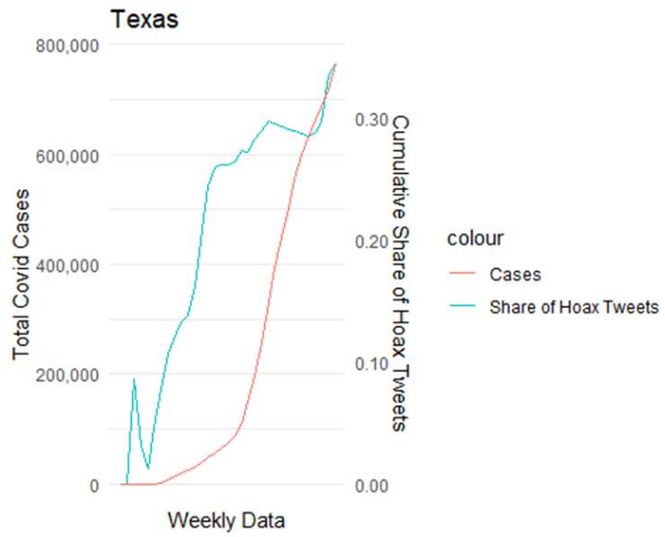
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # 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







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
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     Max
## -294697  -38574  -24113    3608  732032
##
## 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(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

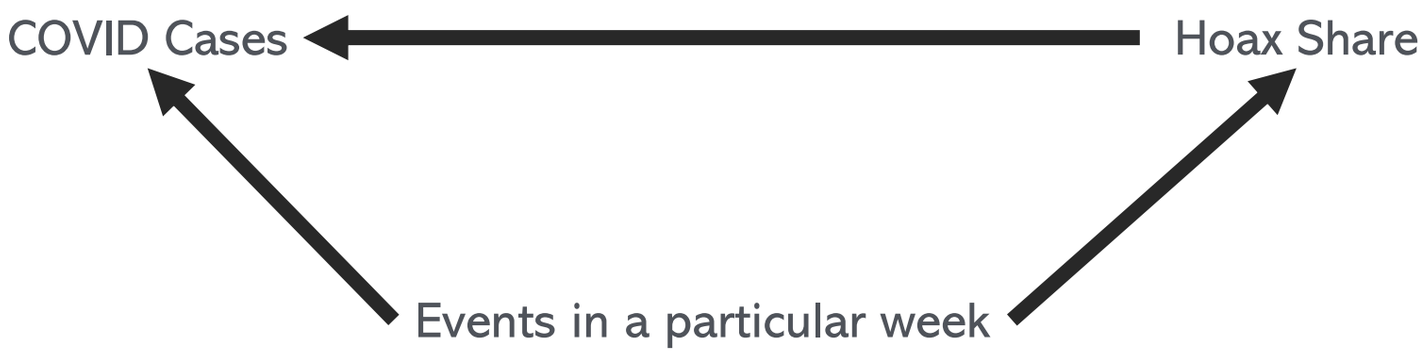
```
lm(cases~cumhoaxsh+factor(week), statsbyweek) %>% summary()

##
## Call:
## lm(formula = cases ~ cumhoaxsh + factor(week), data =
statsbyweek)
##
## Residuals:
##     Min      1Q  Median      3Q     Ma
## -182000  -36269   -9070    5947  66561
##
## Coefficients:
##                           Estimate Std. Error 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
```

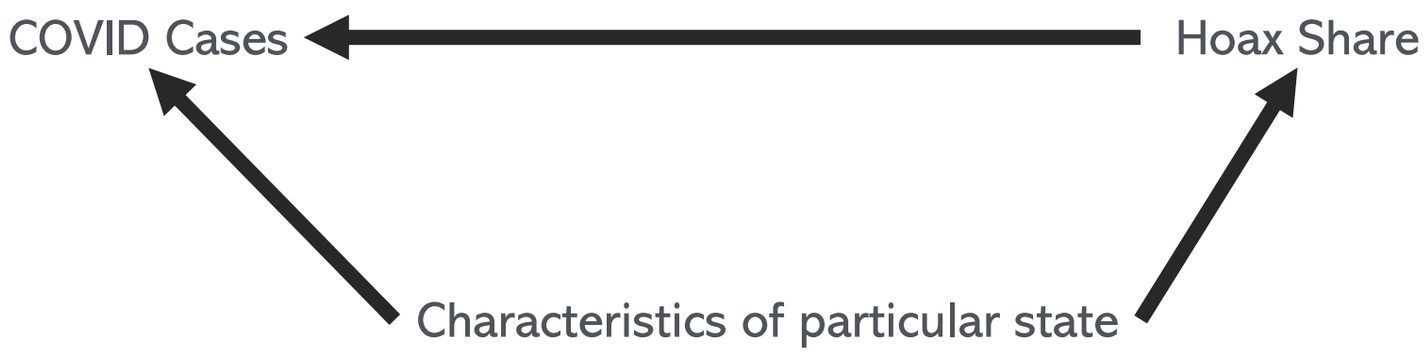
Smaller effect when controlling for time (week) effects

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





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:

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

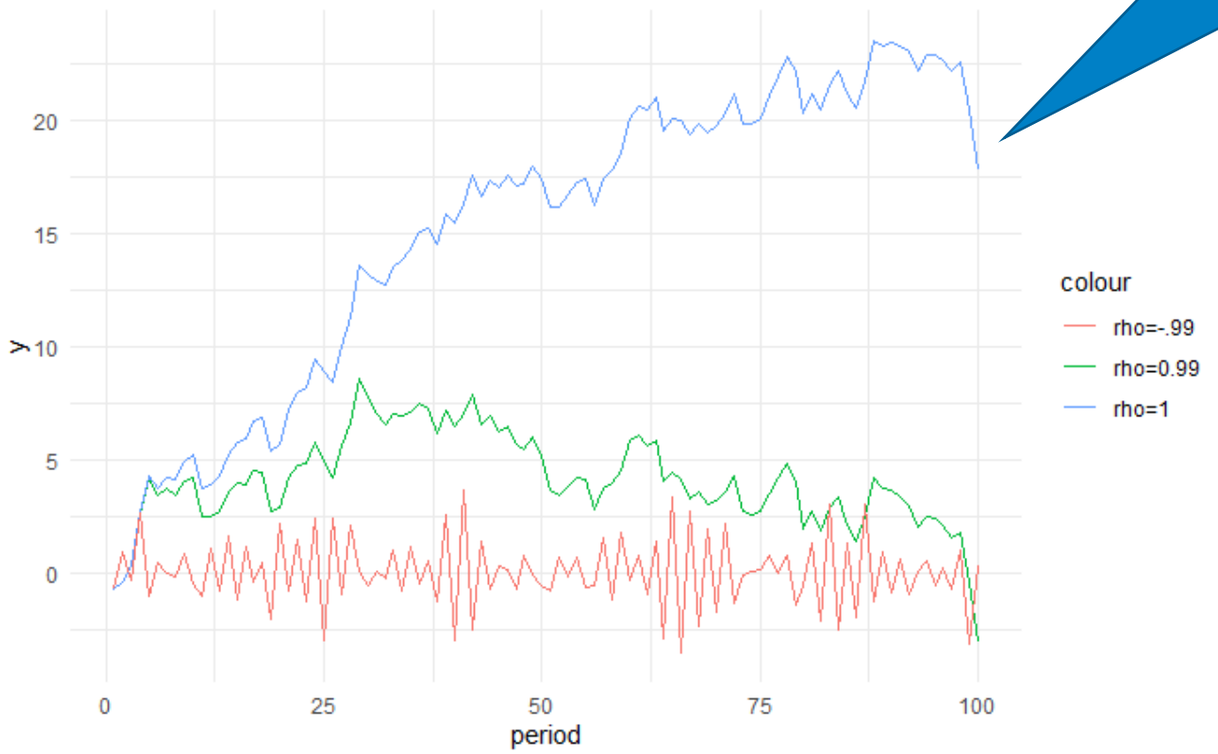
Y_{t-1} becomes the X variable
We can do normal OLS as long as $-1 < \rho < 1$

- 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 $Y = \beta X$ and both Y and X have unit roots you will have a spurious correlation (the unit root becomes the confounder)

- Random Walk

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



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

```
ur.df(diff(df$cases,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
## -0.02330  0.00000  0.00000  0.00000  0.04392
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## z.lag.1    -0.036593   0.007388  -4.953 9.26e-07 ***
## z.diff.lag  0.604696   0.031607  19.132  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.003285 on 666 degrees of freedom
## Multiple R-squared:  0.3567, Adjusted R-squared:  0.3547
## F-statistic: 184.6 on 2 and 666 DF,  p-value: < 2.2e-16
##
##
## Value of test-statistic is: -4.9534
##
## 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 because -4.9534<-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
```

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

Revisiting COVID vs GDP

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


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

```
##
## Call:
## lm(formula = Dlnindex ~ Dcases + t, data = df)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.107e-03 -9.941e-05  4.439e-05  1.487e-04  1.041e-03
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.429e-04  2.415e-05   5.918 5.20e-09 ***
## Dcases      -2.316e-02  7.258e-04 -31.914  < 2e-16 ***
## t            5.269e-07  6.490e-08   8.119 2.28e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual stan      r: 0.000305 on 667 degrees of freedom
##   (1 observat          to missingness)
## Multiple R-sq        d R-squared:  0.6053
## F-statistic:           e: < 2.2e-16
```

100k more cases = 2.3% lower GDP…similar to what we had before….but of course we didn't know that would happen

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

Extra Slides

Other considerations

```
lm(Dlnindex~Dcases+t+Dlockshare,df) %>% summary()
```

```
##
## Call:
## lm(formula = Dlnindex ~ Dcases + t + Dlockshare, data = df)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0011149 -0.0001001  0.0000414  0.0001472  0.0010273
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.401e-04  2.404e-05   5.831 8.61e-09 ***
## Dcases      -2.311e-02  7.221e-04 -32.010  < 2e-16 ***
## t            5.400e-07  6.471e-08   8.345 4.10e-16 ***
## Dlockshare  -1.253e-05  4.354e-06  -2.878  0.00412 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standa          0.0003034 on 666 degrees of freedom
##   (1 observatio            to missingness)
## Multiple R-squa            ted R-squared:  0.6095
## F-statistic: 34            value: < 2.2e-16
```

- If 100% of US population go into lockdown GDP goes down by -0.138% (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

$$\beta_1 + \beta_2 = 1$$

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

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

More lags and trend?

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

More lags

```
lm(Dlnindex~dplyr::lag(Dlnindex)+dplyr::lag(Dlnindex,2)+Dcases+dplyr::lag(Dcases)+dplyr::lag(Dcases,2)+t+Dlockshar
e+dplyr::lag(Dlockshare)+dplyr::lag(Dlockshare,2),df) %>% summary()
```

```
##
## Call:
## lm(formula = Dlnindex ~ dplyr::lag(Dlnindex) + dplyr::lag(Dlnindex,
##     2) + Dcases + dplyr::lag(Dcases) + dplyr::lag(Dcases, 2) +
##     t + Dlockshare + dplyr::lag(Dlockshare) + dplyr::lag(Dlockshare,
##     2), data = df)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.264e-04 -3.238e-05 -2.604e-06  3.437e-05  1.893e-04
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.374e-06  4.565e-06   0.301   0.7634
## dplyr::lag(Dlnindex)        7.911e-01  3.793e-02  20.855  < 2e-16 ***
## dplyr::lag(Dlnindex, 2)     1.854e-01  3.743e-02   4.954 9.27e-07 ***
## Dcases                     -7.495e-04  1.100e-03  -0.681   0.4960
## dplyr::lag(Dcases)         -3.192e-03  1.444e-03  -2.210   0.0275 *
## dplyr::lag(Dcases, 2)       3.703e-03  7.344e-04   5.043 5.94e-07 ***
## t                           2.186e-08  1.270e-08   1.721   0.0857 .
## Dlockshare                 -1.036e-05  8.940e-07 -11.586  < 2e-16 ***
## dplyr::lag(Dlockshare)     -9.179e-06  1.167e-06  -7.867 1.49e-14 ***
## dplyr::lag(Dlockshare, 2) -3.171e-06  1.449e-06  -2.189   0.0290 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.585e-05 on 658 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.987,  Adjusted R-squared:  0.9868
## F-statistic:  5546 on 9 and 658 DF,  p-value: < 2.2e-16
```

Further reading

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