Time Series

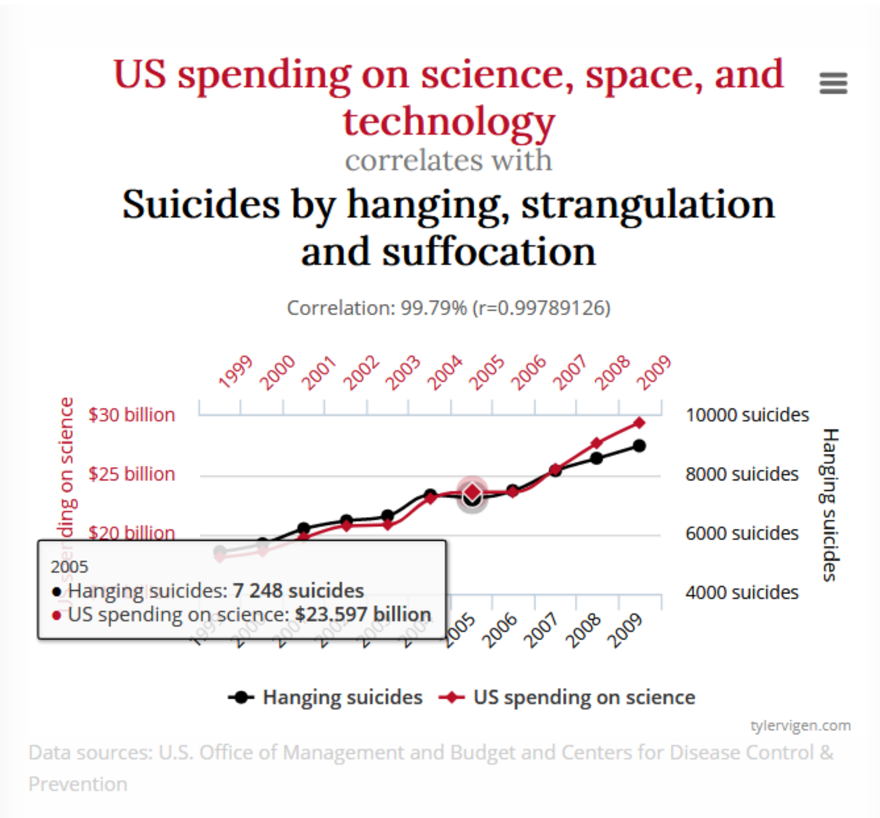
To infinity and beyond….

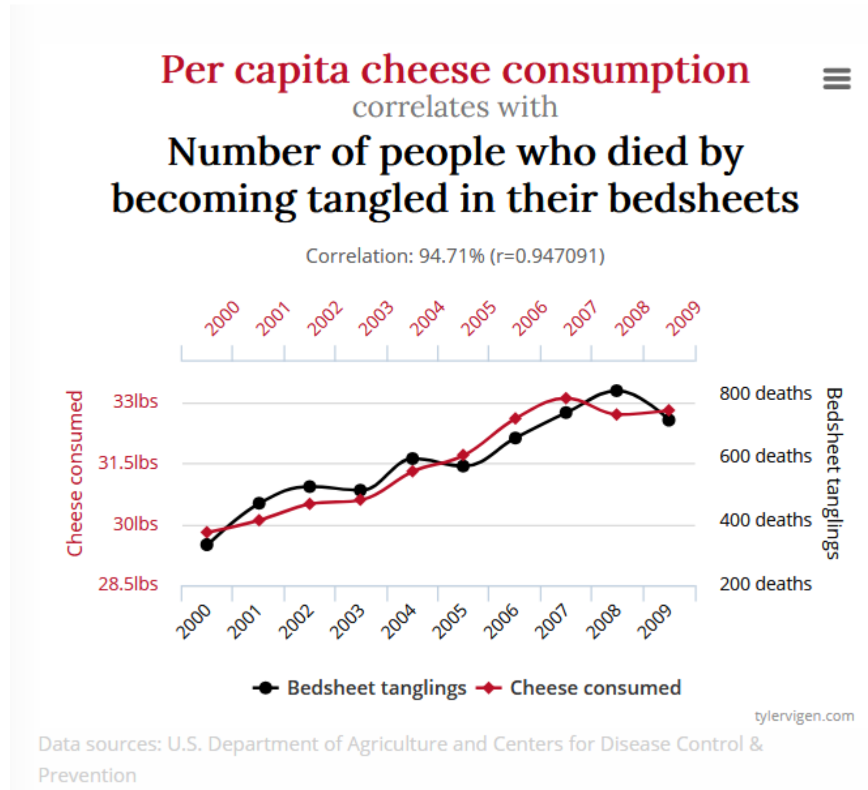
by Ralf Martin (r.martin@imperial.ac.uk)

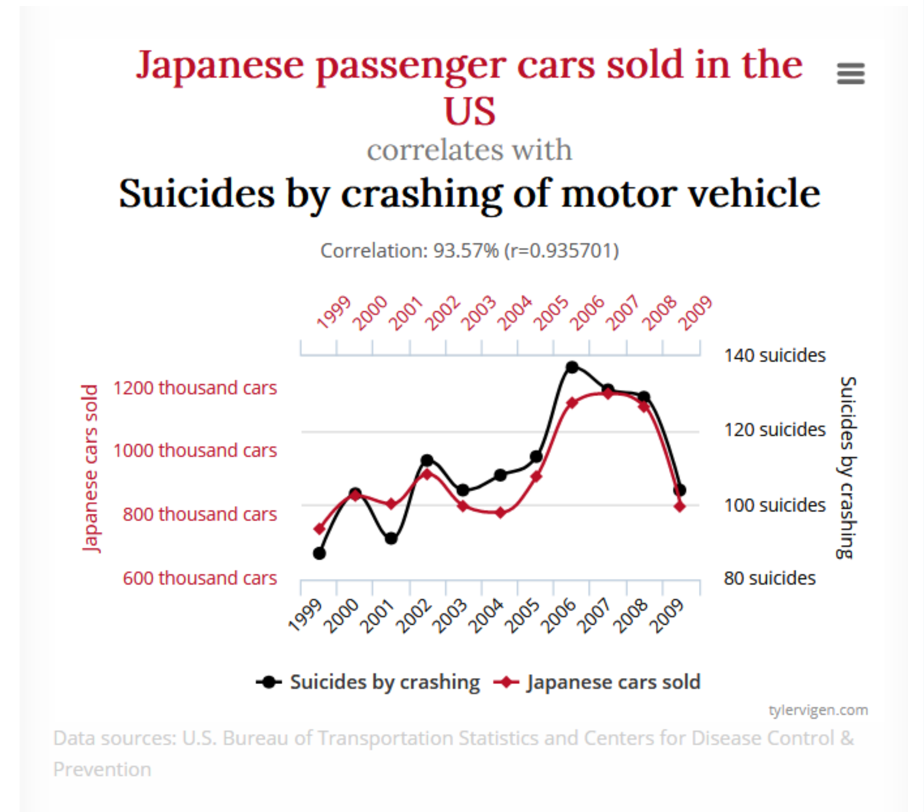
Objectives of this lecture

- 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?



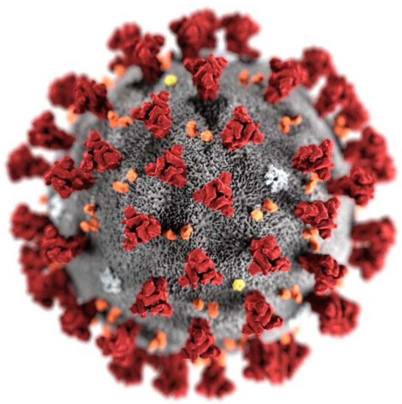




Can you spot the problem?

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

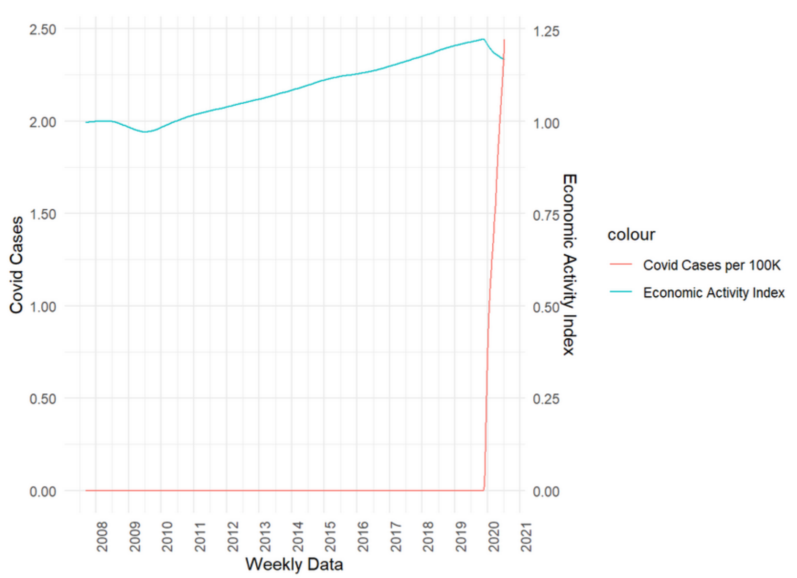
COVID vs GDP



```
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?



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  0.05[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   es:  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))
```

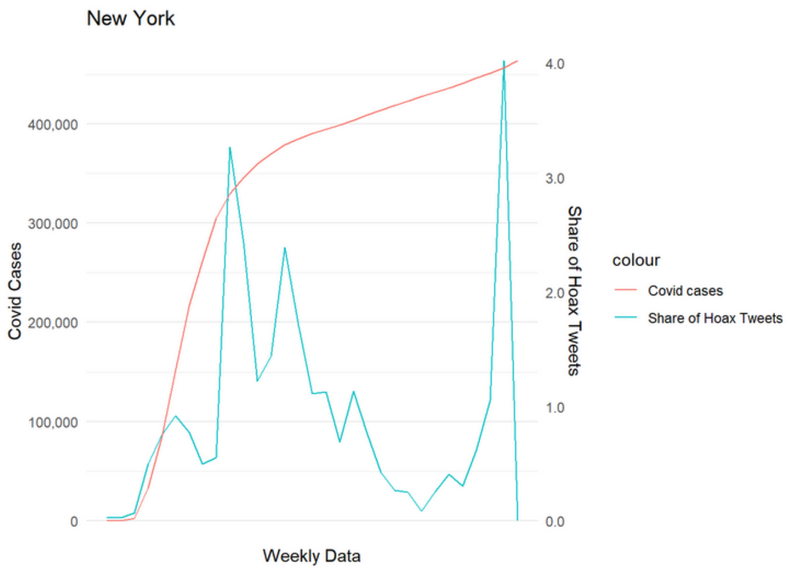
```
## # A tibble: 6 x 9
## # Groups:   state [1]
##   state   week        hoax tweets cases deaths hoaxsh Dcases Ddeaths
##   <chr>   <date>     <int>  <int> <int>  <int>  <dbl>  <int>   <int>
## 1 Alabama 2020-03-15     4   1503    51      0  0.266     NA      NA
## 2 Alabama 2020-03-22    62   4198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

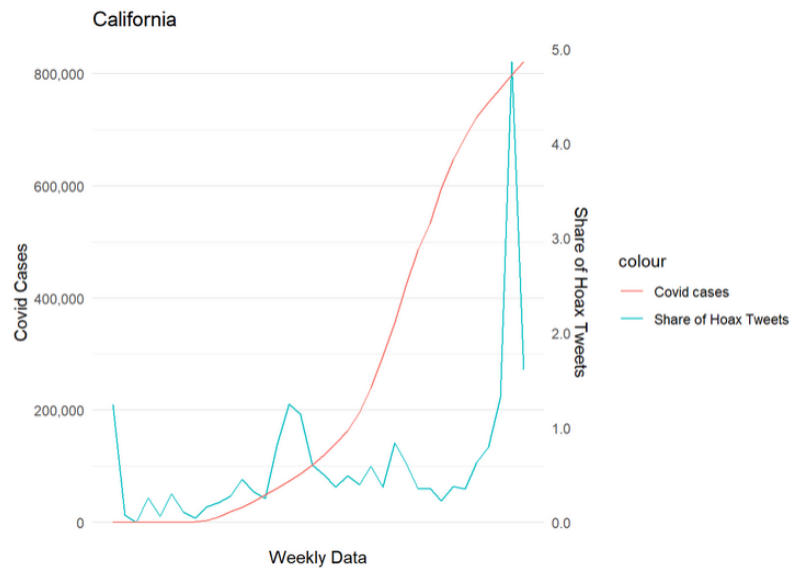
```
statsbyweek %>% group_by(state) %>% summarise(n())
```

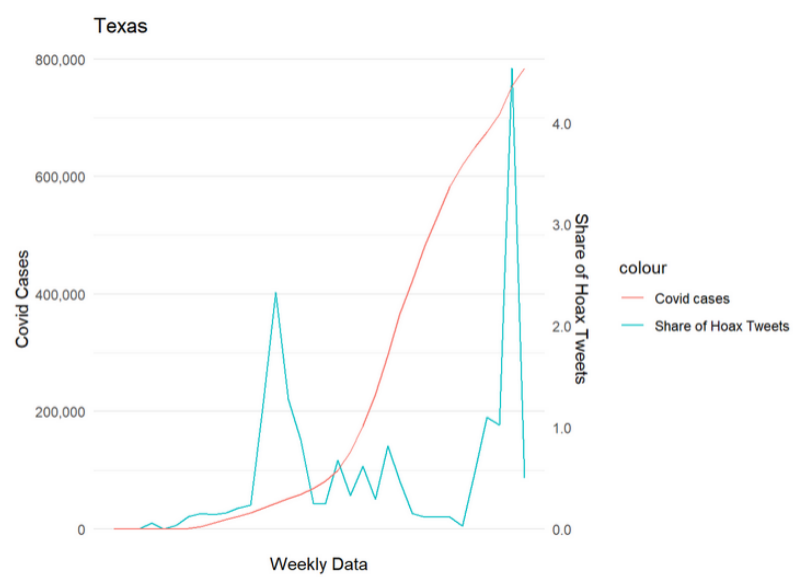
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # 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 example

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

```
##
## Call:
## lm(formula = cases ~ hoaxsh, data = statsbywe
##
## Residuals:
##     Min      1Q  Median      3Q     M
## -189328  -50914  -40048    7176   7
##
## Coefficients:
##             Estimate Std.  rror t value Pr(>|t|)
## (Intercept)    50929       3108  16.388  < 2e-16 ***
## hoaxsh         11555       2380   4.855 1.33e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 108700 on 1544 degrees of freedom
## Multiple R-squared:  0.01504,   Adjusted R-squared:  0.0144
## F-statistic: 23.57 on 1 and 1544 DF,  p-value: 1.326e-06
```

Hoax share up by 1 percentage point means 11555 more cases

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

```
##
## Call:
## lm(formula = cases ~ hoaxsh + factor(wee
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -199861  -37318   -9820    1098  668461
##
## Coefficients:
##                          Estimate Std.  or t value Pr(>|t|)
## (Intercept)                 1.00   956.85   0.000  0.99999
## hoaxsh                   7865.20  2593.18   3.033  0.00246 **
## factor(week)2020-01-26  -2439.83 111758.05  -0.022  0.98259
## factor(week)2020-02-02    -95.32 107965.74  -0.001  0.99930
## factor(week)2020-02-09      1.00 106858.37   0.000  0.99999
## factor(week)2020-02-16   -254.24 106020.28  -0.002  0.99809
## factor(week)2020-02-23    -50.25 105363.77   0.000  0.99962
## factor(week)2020-03-01  -1014.28 102855.30  -0.010  0.99213
## factor(week)2020-03-08    -70.35 101086.35  -0.001  0.99944
## factor(week)2020-03-15   -133.49 100951.52  -0.001  0.99895
```

Smaller effect when controlling for time (week) effects

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

```
##
## Call:
## lm(formula = cases ~ hoaxsh + +factor(state) + factor(week),
##     data = statsbyweek)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -264367  -23041     593   22221  456332
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -1002     70567  -0.014 0.988669
## hoaxsh                       3788      1863   2.033 0.042192 *
## factor(state)Alaska        -52626     17974  -2.928 0.003465 **
## factor(state)Arizona        41125     17193   2.392 0.016885 *
## factor(state)Arkan         -24748     17843  -1.387 0.165651
## factor(state)Califo        223775     17186  13.021  < 2e-16 ***
##                                       17833  -1.107 0.268403
```

Also controlling for state

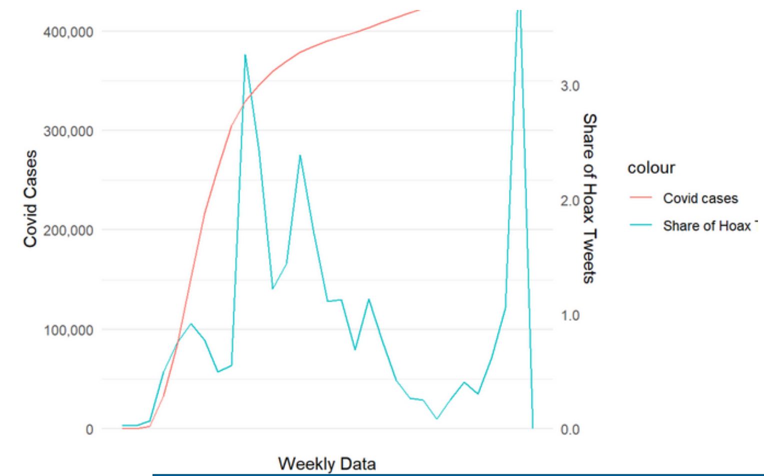
```
library(plm)
plm(cases~hoaxsh+factor(week)+factor(state),statsbyweek
    index=c("state","week"),
    model="within",
    effect="twoways") %>% summary()
```

Alternative command to include cross sectional and time effects in panel data Substantially more efficient with large datasets (many cross sectional units)

```
## Twoways effects Within Model
##
## Call:
## plm(formula = cases ~ hoaxsh + factor(wee
##     data = statsbyweek, effect = "twoways", model = "within",
##     index = c("state", "week"))
##
## Unbalanced Panel: n = 50, T = 29-37, N = 1546
##
## Residuals:
##      Min.    1st Qu.     Median    3rd Qu.      Max.
## -264367.42  -23040.53    592.79   22221.48  456331.70
##
## Coefficients:
##        Estimate Std. Error t-value Pr(>|t|)
## hoaxsh   3788.3     1863.0  2.0334  0.04219 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    6.8535e+12
## Residual Sum of Squares: 6.8341e+12
## R-Squared:      0.0028259
## Adj. R-Squared: -0.055952
## F-statistic: 4.13474 on 1 and 1459 DF, p-value: 0.042192
```

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:

- $Y_t = \beta_0 + \beta_1 Y_{t-1} + \epsilon_t$

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:

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

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

```r
library(urca)
```

```
## Warning: package 'urca' was built under R version 4.0.2
```

```r
ur.df(df$cases,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
## -0.02151  0.00000  0.00000  0.00000  0.07106
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## z.lag.1    0.0004516  0.0006910   0.654    0.514
## z.diff.lag 0.9805725  0.0133827  73.272   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.004085 on 667 degrees of freedom
## Multiple R-squared:  0.9468, Adjusted R-squared:  0.9467
## F-statistic:  5938 on 2 and 667 DF,  p-value: < 2.2e-16
##
##
## Value of test-statistic is: 0.6536
##
## Critical values for test statistics:
##       1pct  5pct 10pct
## tau1 -2.58 -1.95 -1.62
```

We cannot reject unit root becase 0.653>-1.95

```r
ur.df(df$lnindex,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
## -8.343e-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 becase -0.7971>-1.95

Getting rid of unit roots

```r
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

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 observati              o missingness)
## Multiple R-squa              ted R-squared:  0.6095
## F-statistic: 3                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

$$\beta_1 + \beta_2 = 1$$

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

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

More lags and trend?

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

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]