COVID-19 Data Analysis & Prediction:

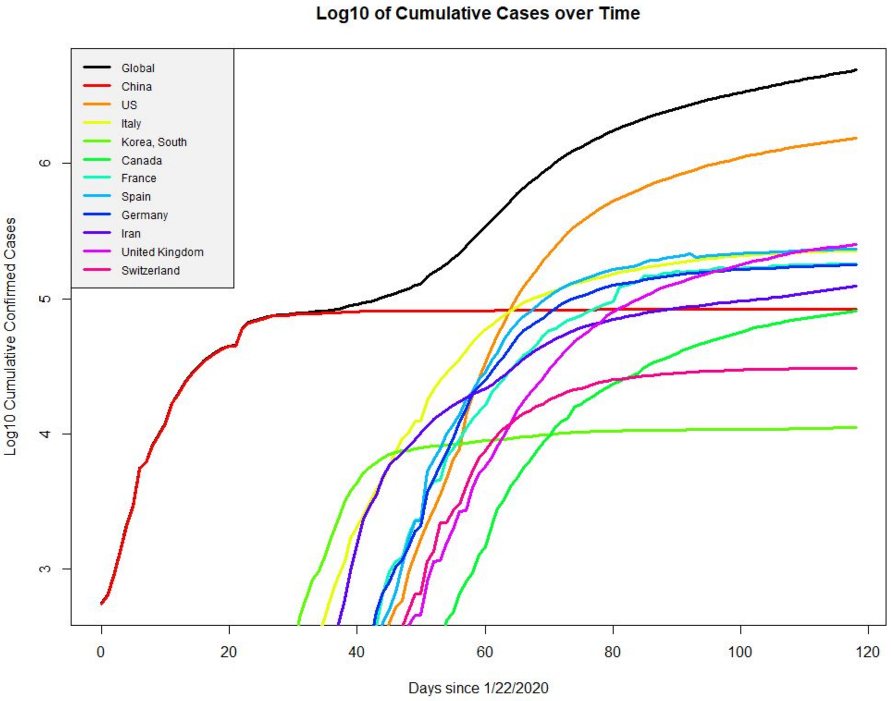
A systematic exploration of COVID-19's toll on the world population, with efforts made both to predict case counts and to ascertain salient demographic predictors of death and disease transmission rates.

By Clayton Bass, Thomas Brown, Kayshav Prakash,
Bruce Zou, Allison Tong

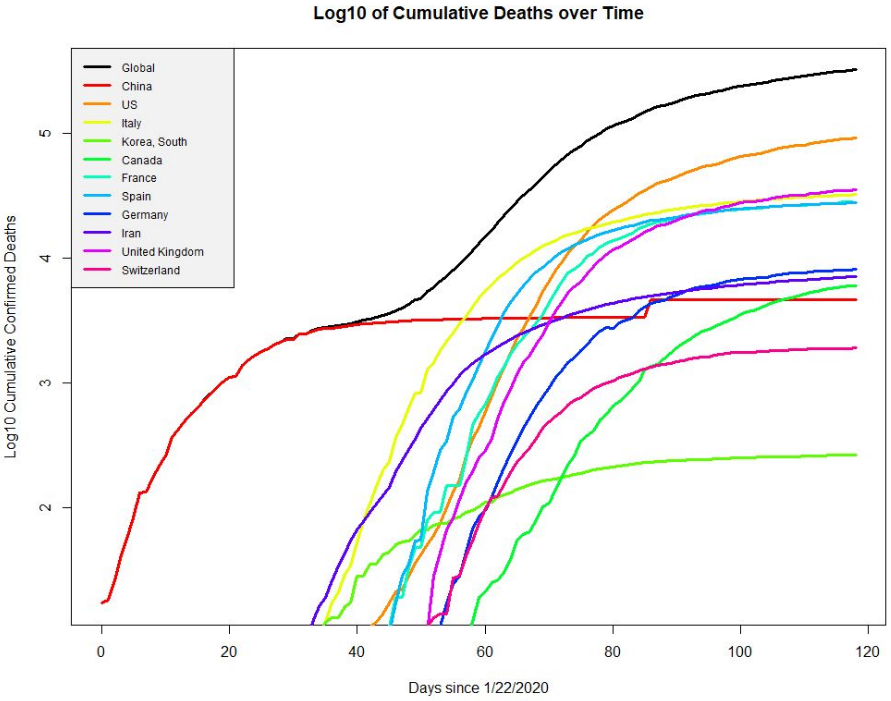
Professor Demidenko's 2020
MATH 70 Class

Section 1: Big-picture analysis and trends.

Log10 of deaths and cases over time



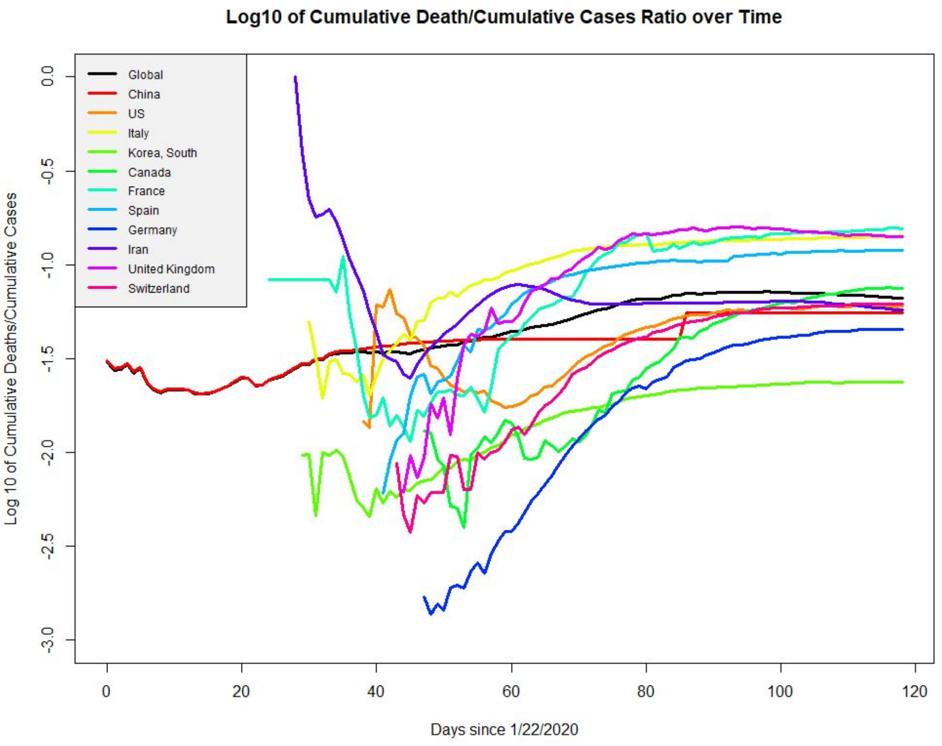
Log10 of Cumulative Cases over Time



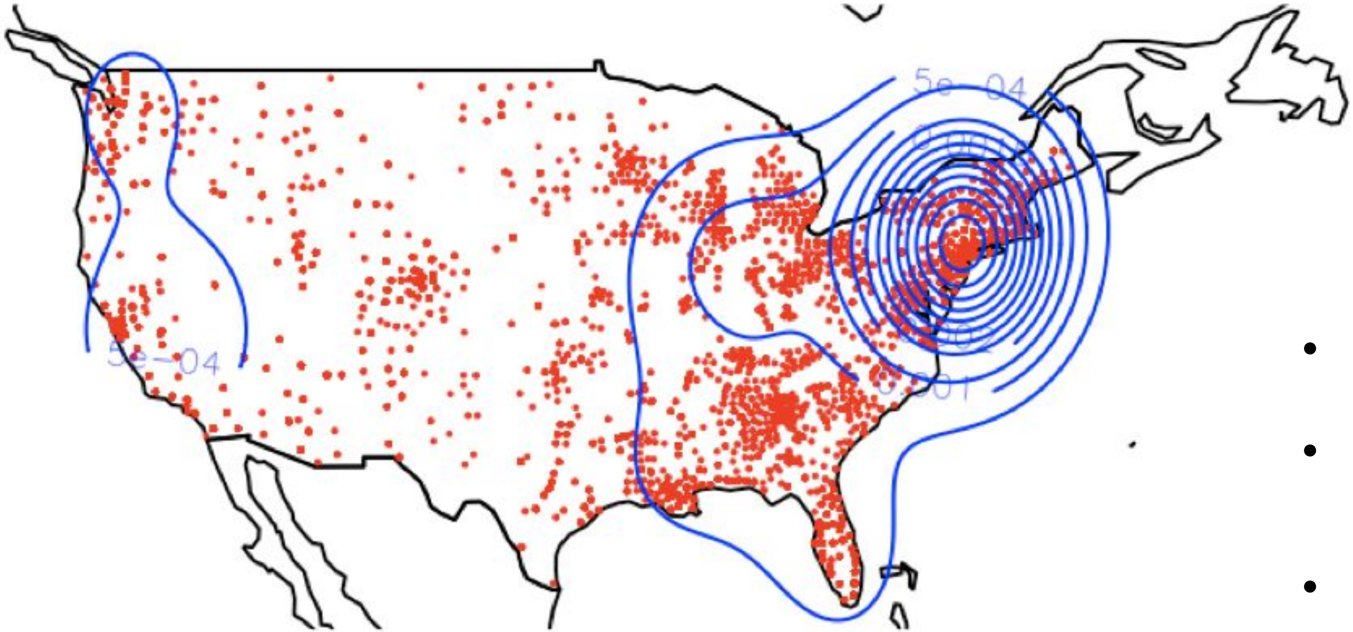
Log10 of Cumulative Deaths over Time

Data Source: JHU Time Series (https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data)

Log10 of Deaths/Cases over time



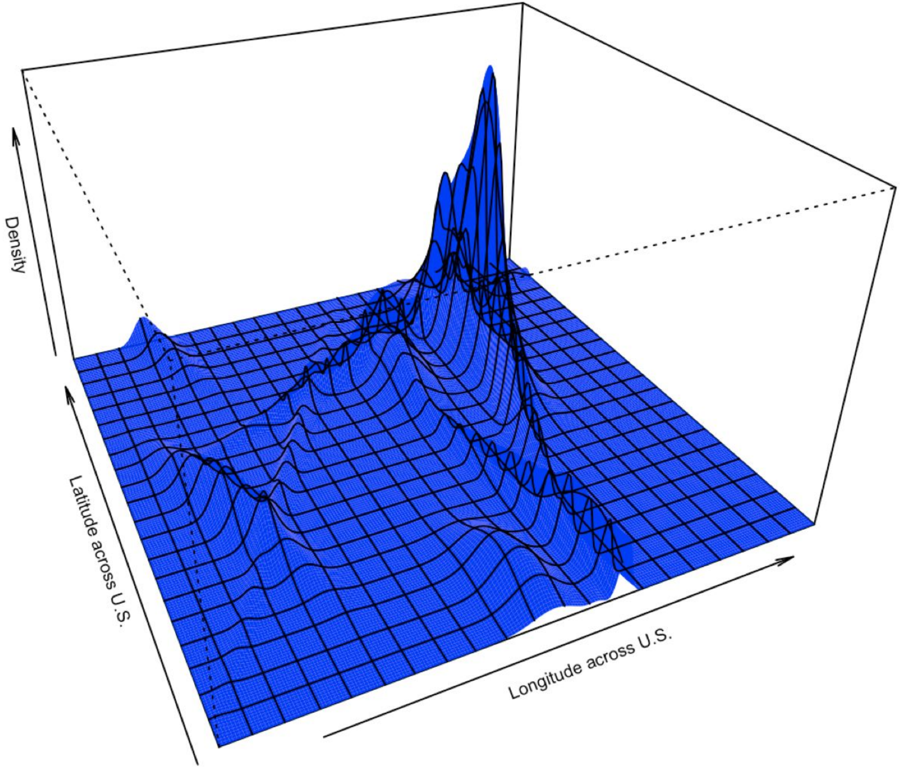
Bivariate density estimate & contour plot for U.S. COVID-19 case density; optimal cross-validation bandwidth h = 3.404



- Coordinate data provided up to April (peak of COVID-19)
    - Coordinates of each recorded case seem to indicate county belonging (*not* exact case location)
        - Red dots are essentially just affected counties
        - Cases look sparser than they truly are
    - Bivariate density, however, *does* reflect concentration of individual cases
- NYC is *by far* the epicenter, dominating the bivariate density
- The Midwest has vanishing density values
    - Comparatively few cases
- California demonstrates elevated case density

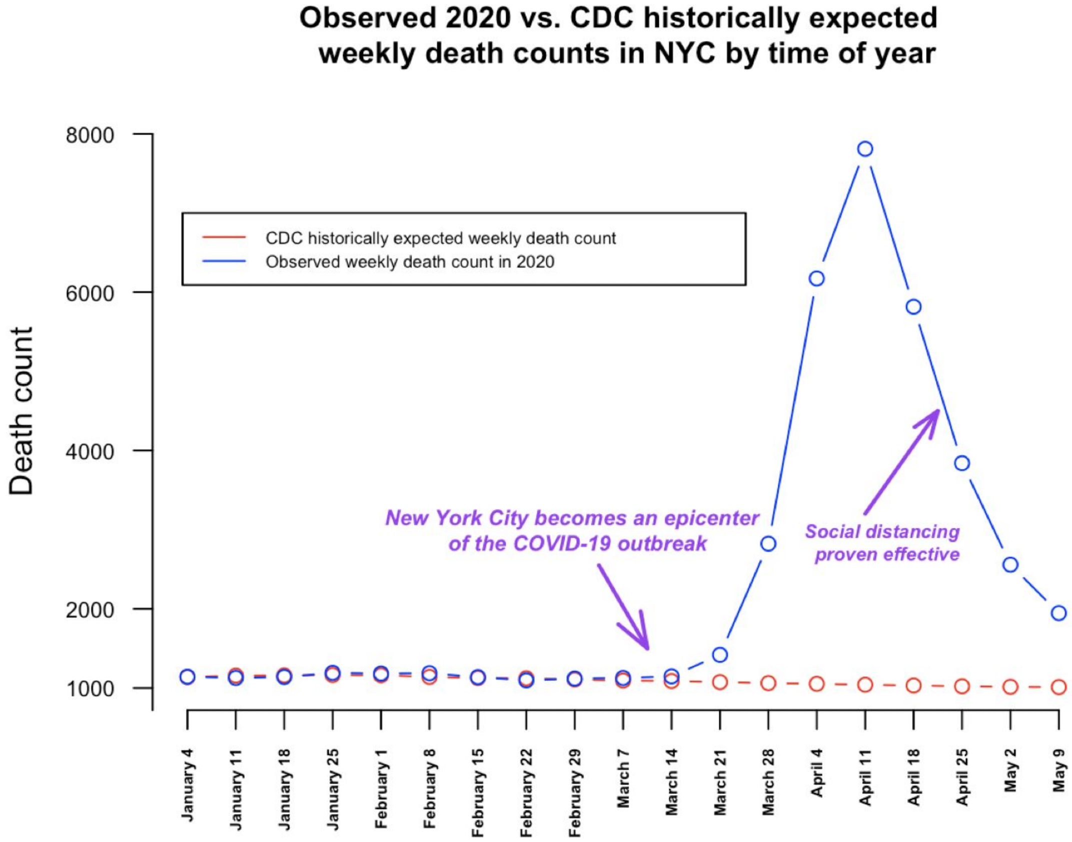
The big picture: bivariate case density (contour plot) in the United States, with optimal bandwidth estimated via cross-validation. Source for rworldmap R package guide: http://www.milanor.net/blog/maps-in-r-plotting-data-points-on-a-map/. Source of data: https://github.com/beoutbreakprepared/nCoV2019/blob/master/latest_data/latestdata.tar.gz.

mclust *persp* plot of case density across the U.S. at peak of spread rate (April)



- Reiterates but visually explains NYC as the major epicenter at the peak of the COVID-19 pandemic
- The shape of the continental U.S. (especially Florida) can be inferred from the plot
- A string of cases throughout California

Three-dimensional *persp* plot of cases across the U.S., thanks to the mclust package with *type* = "*persp*" in the *plot* call. (Source for mclust code formatting: https://cran.r-project.org/web/packages/mclust/vignettes/mclust.html, in the "Multivariate" section. Data source same as previous slide.)



- Death counts consistent with historical expectation until March 21
    - Over 8x as many deaths as usual in mid-April
    - CDC (per data in description below) places **upper bound of normal count** (natural variation) at **1121 deaths** for week of April 11...
- Peak collateral damage (elevated death rate) in the week of April 11
- Social distancing proves effective by the end of April

Hot spot case study: NYC collateral damage estimates, from January 4 until May 9 (updated somewhat irregularly). **Source**: CDC, at https://data.cdc.gov/NCHS/Excess-Deaths-Associated-with-COVID-19/xkkf-xrst.

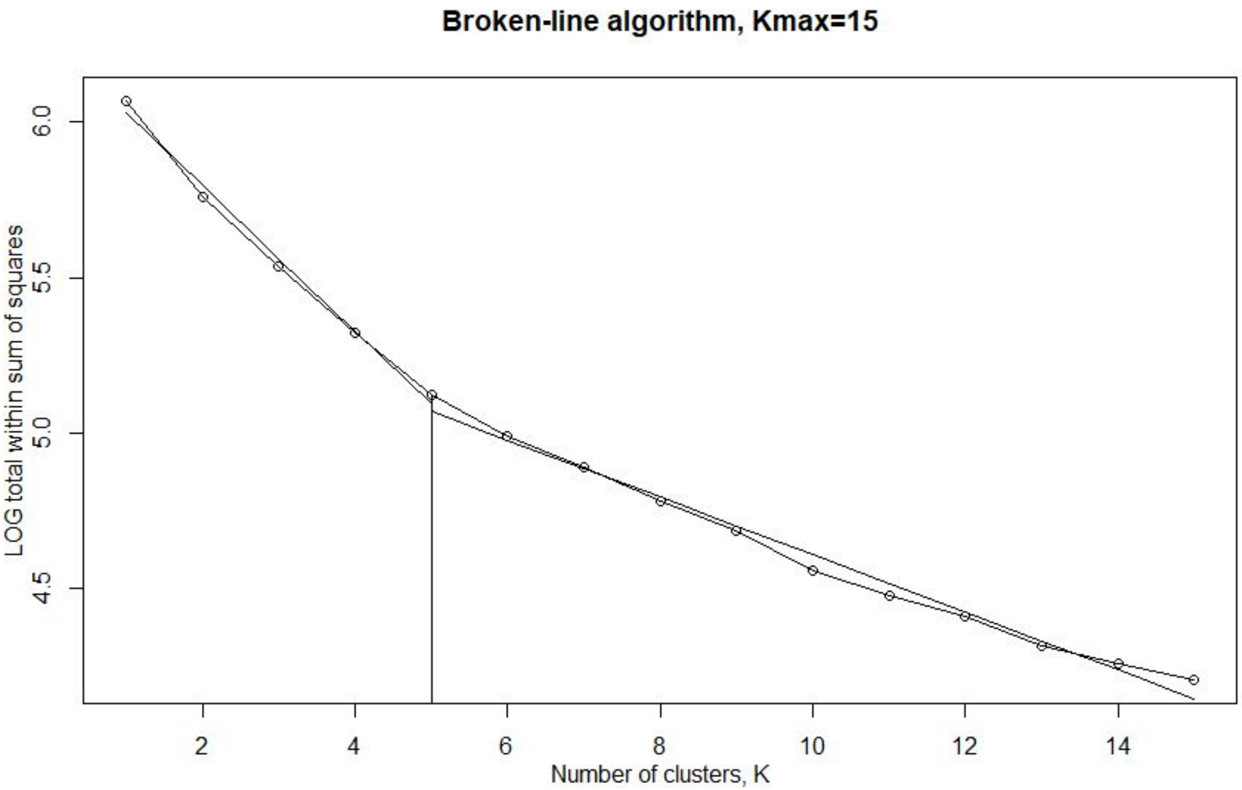
Cluster analysis of case data by country (Bruce)

Features:

- INFORM Covid-19 Hazard & Exposure Risk (0-10, Higher is worse)
- INFORM Covid-19 Vulnerability Risk (0-10, Higher is worse)
- INFORM Covid-19 Lack of Coping Capacity Risk (0-10, Higher is worse)
- Early Spread Days = Days between outbreak (5 cases/million) and 100 cases/million
- Testing Delay = Days between outbreak and reaching 5 tests/thousand
- Total cases/million
- Total deaths/million
- Total tests/thousand

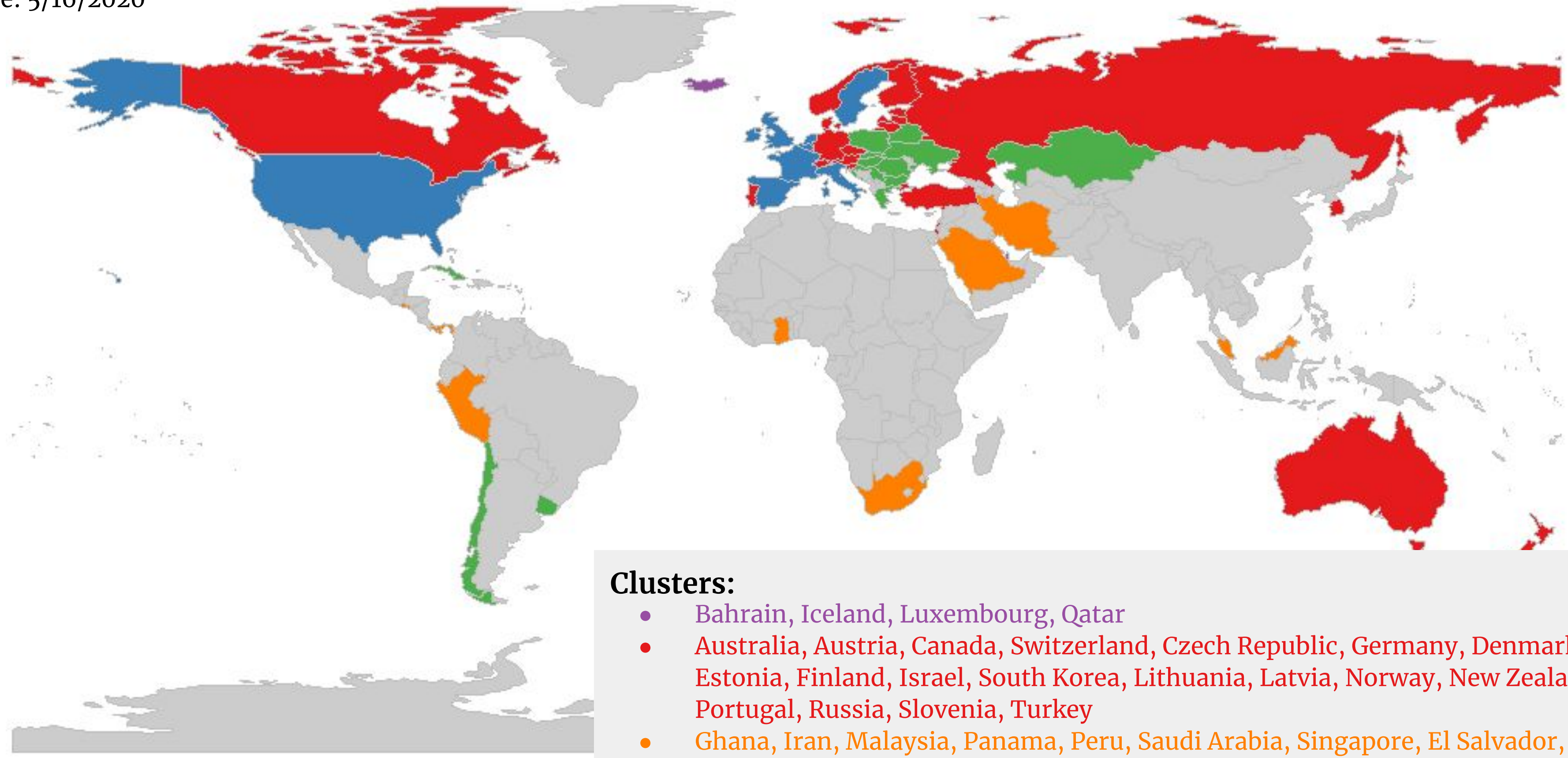
Cases, Deaths, and Testing data (Our World in Data) from 5/16/2020

Data Sources: https://ourworldindata.org/mortality-risk-covid & https://drmkc.jrc.ec.europa.eu/inform-index/INFORM-Epidemic



Optimal number of clusters, K = 5

Date: 5/16/2020



K-means Clusters Map, K=5

Clusters:

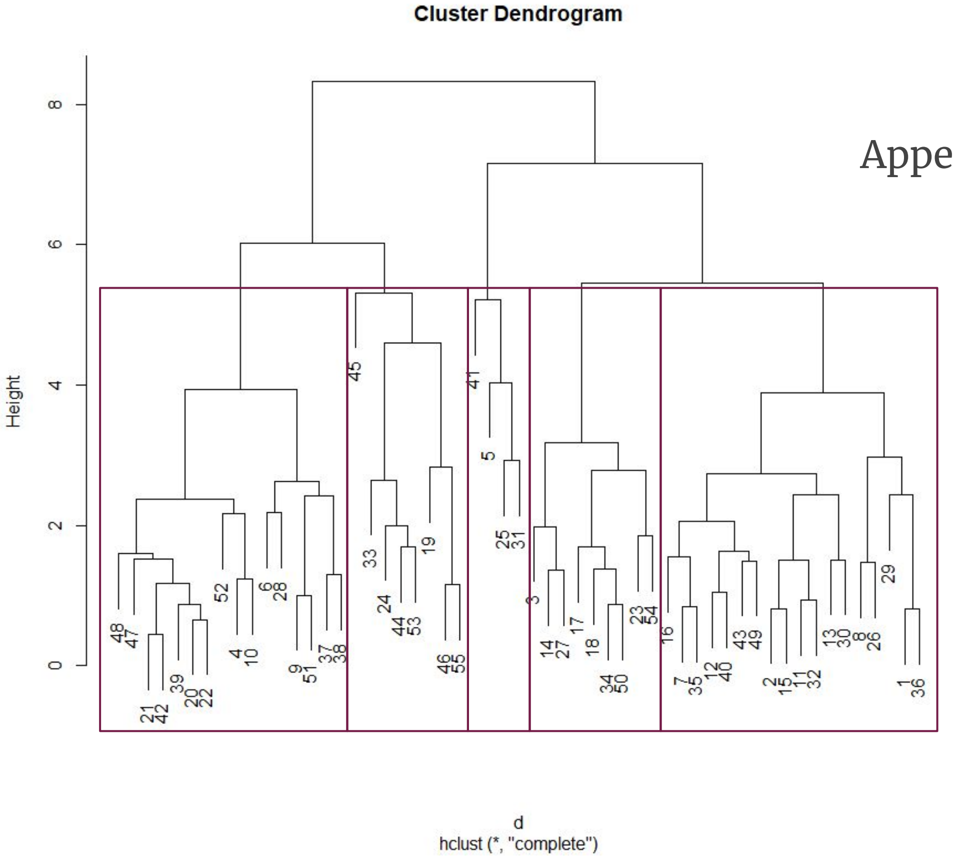
- Bahrain, Iceland, Luxembourg, Qatar
- Australia, Austria, Canada, Switzerland, Czech Republic, Germany, Denmark, Estonia, Finland, Israel, South Korea, Lithuania, Latvia, Norway, New Zealand, Portugal, Russia, Slovenia, Turkey
- Ghana, Iran, Malaysia, Panama, Peru, Saudi Arabia, Singapore, El Salvador, South Africa
- Bulgaria, Belarus, Chile, Cuba, Greece, Croatia, Hungary, Kazakhstan, Poland, Romania, Serbia, Slovakia, Ukraine, Uruguay
- Belgium, Spain, France, United Kingdom, Ireland, Italy, Netherlands, Sweden, United States

K-Means Analysis: Cluster Means

	Hazard & Exposure Risk	Vulnerability Risk	Lack of Coping Capacity Risk	Early Spread Days	Testing Delay (days)	Cases /Million	Deaths /Million	Tests /Thousand
Purple	4.75	4.95	2.0625	8.25	16.75	6407.401	51.839	112.6705
Red	2.684211	6.594737	1.115789	10.789474	16.473684	1406.8723	55.979526	42.465263
Orange	3.988889	4.288889	3.033333	27.333333	44.333333	1436.0623	26.750778	13.338667
Green	2.535714	6.821429	3.235714	20.357143	34.785714	764.117357	19.916786	16.315071
Blue	2.988889	6.822222	1.038889	10.888889	29.222222	3739.51189	452.00089	32.689333

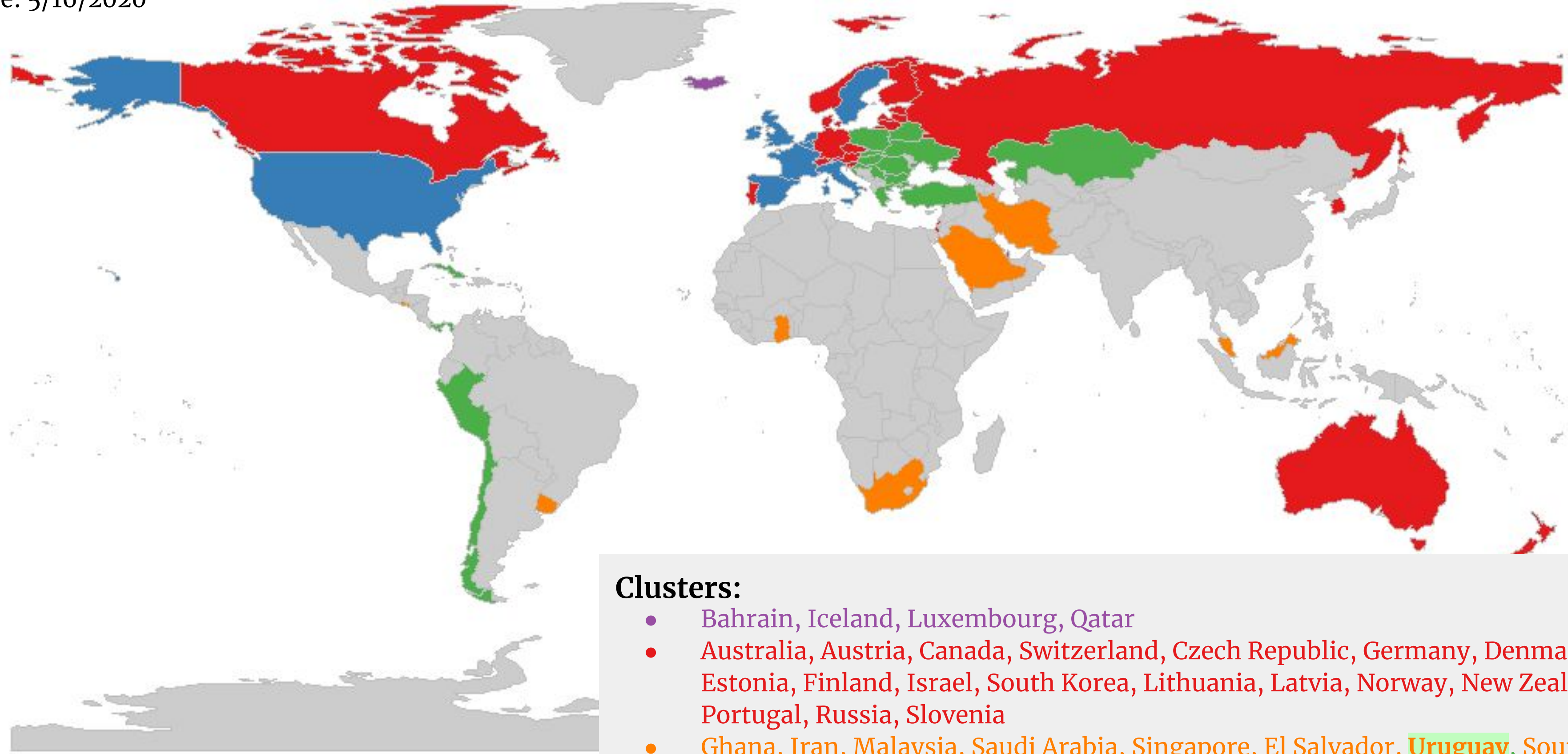
- Red (Canada, S. Korea) vs Blue (US, Italy, Spain), similar risk scores and early spread, but big difference in testing delay, cases/millions and deaths/millions
- Orange (Iran, Saudi Arabia, Singapore) and Green (Chile, Greece, Ukraine) both have slower early spread and lower deaths/million despite having low tests/thousand
- Purple (Iceland, Luxembourg) had fast early spread, high number of cases/million, and high tests/thousand

Hierarchical Clustering Dendrogram



Appears to have 5 clusters

Date: 5/16/2020

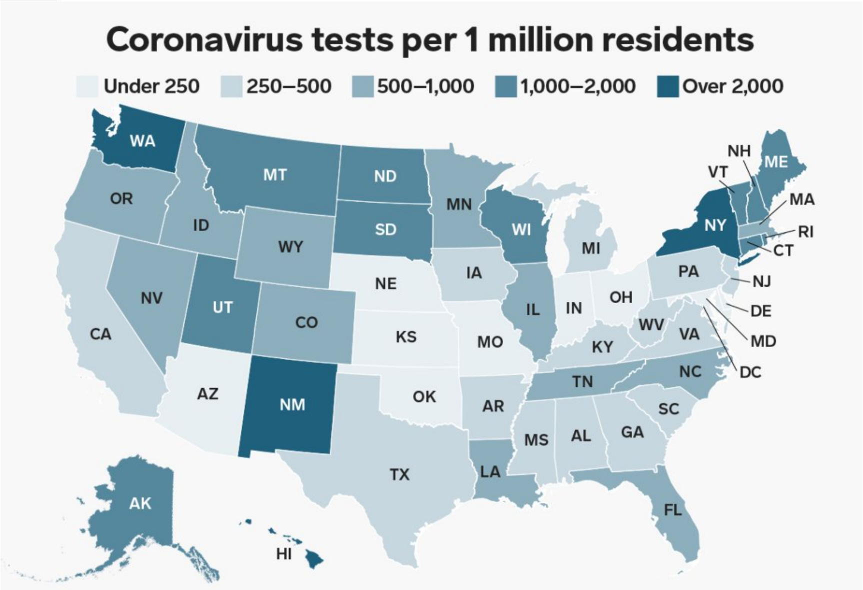


Clusters:

- ● Bahrain, Iceland, Luxembourg, Qatar
- ● Australia, Austria, Canada, Switzerland, Czech Republic, Germany, Denmark, Estonia, Finland, Israel, South Korea, Lithuania, Latvia, Norway, New Zealand, Portugal, Russia, Slovenia
- ● Ghana, Iran, Malaysia, Saudi Arabia, Singapore, El Salvador, **Uruguay**, South Africa
- ● Bulgaria, Belarus, Chile, Cuba, Greece, Croatia, Hungary, Kazakhstan, **Panama**, **Peru**, Poland, Romania, Serbia, Slovakia, , **Turkey**, Ukraine
- ● Belgium, Spain, France, United Kingdom, Ireland, Italy, Netherlands, Sweden, United States

Hierarchical Clusters Map, K=5

High Resolution Data for tracking variables related to COVID deaths



Source: Business Insider/COVID Tracking Project 3/23
https://www.businessinsider.com/map-us-states-coronavirus-case-totals-cases-per-capita-tests-2020-5

Barriers

- Nationally, the highest resolution data provided is at the county level. This data, while allowing for very broad trends to be studied isn’t ideal as counties can vary immensely in size/diversity (e.g. LA County)
- Not all US localities are reporting both deaths and cases by county in a consistent way. States with many counties tend to have much missing data.
- Death rates themselves are not an entirely reliable measure when used nationally. States vary greatly in the amount of tests they are giving per-capita and this informs the case count.
- COVID has hit regions at different times and to different regions. Because of this, comparing deaths/capita has limited meaning

Workarounds

- Focus on a specific region, with consistent policies across the board when it comes to testing/reporting data for COVID-19

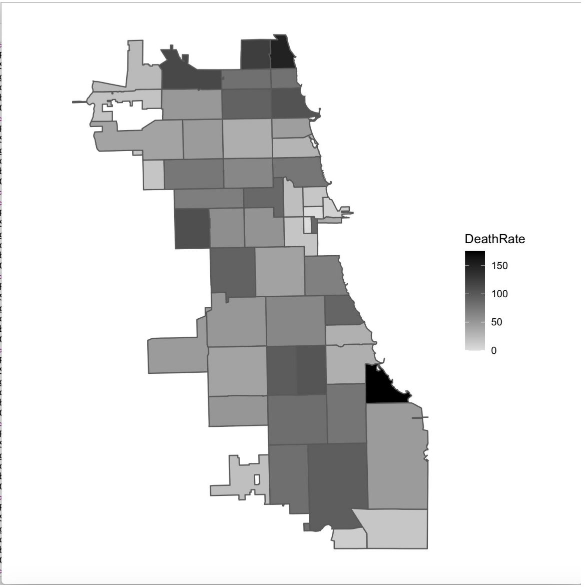
- Use publicly available municipal data combined with case counts in the areas to create predictors of the mortality from COVID.

- When it comes to reporting, only a few metro areas are reporting beyond the county-level (The places that are going further are reported by ZIP Code)

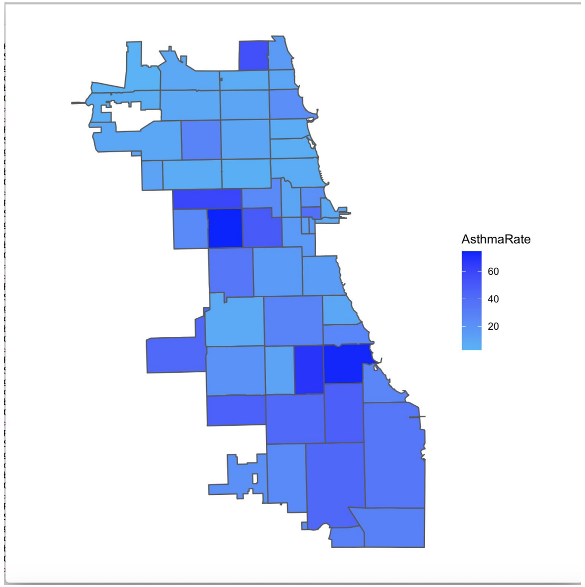
- In addition, the number of health measures publicly available at that same level is quite limited.

Cook County, IL: Health Predictors by ZIP Code

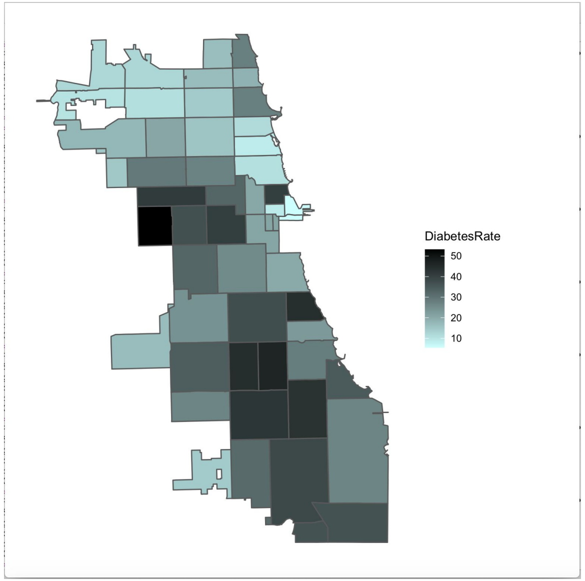
Covid 19 Deaths / 100,000 ppl



Asthma Hospitalization / 10,000 ppl

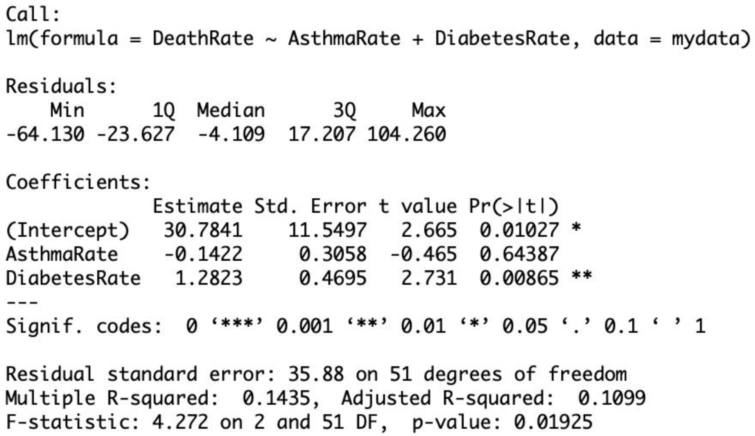


Diabetes Hospitalizations/ 10,000 ppl



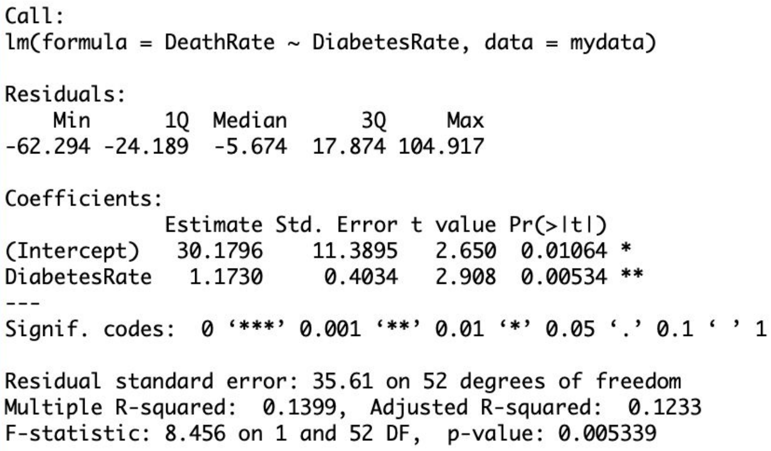
Sources: Chicago Dept. Of Health (COVID) & City of Chicago
https://data.cityofchicago.org/Health-Human-Services/COVID-19-Cases-Tests-and-Deaths-by-ZIP-Code/yhhz-zm2v
https://healthdata.gov/dataset/public-health-statistics-asthma-hospitalizations-chicago-year-2000-2011

Linear Model



```
Call:
lm(formula = DeathRate ~ AsthmaRate + DiabetesRate, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-64.130 -23.627  -4.109  17.207 104.260 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept)   30.7841    11.5497   2.665  0.01027 * 
AsthmaRate    -0.1422     0.3058  -0.465  0.64387   
DiabetesRate   1.2823     0.4695   2.731  0.00865 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 35.88 on 51 degrees of freedom
Multiple R-squared:  0.1435,	Adjusted R-squared:  0.1099 
F-statistic: 4.272 on 2 and 51 DF,  p-value: 0.01925
```

Non-parsimonious Model



```
Call:
lm(formula = DeathRate ~ DiabetesRate, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-62.294 -24.189  -5.674  17.874 104.917 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept)   30.1796    11.3895   2.650  0.01064 * 
DiabetesRate   1.1730     0.4034   2.908  0.00534 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 35.61 on 52 degrees of freedom
Multiple R-squared:  0.1399,	Adjusted R-squared:  0.1233 
F-statistic: 8.456 on 1 and 52 DF,  p-value: 0.005339
```

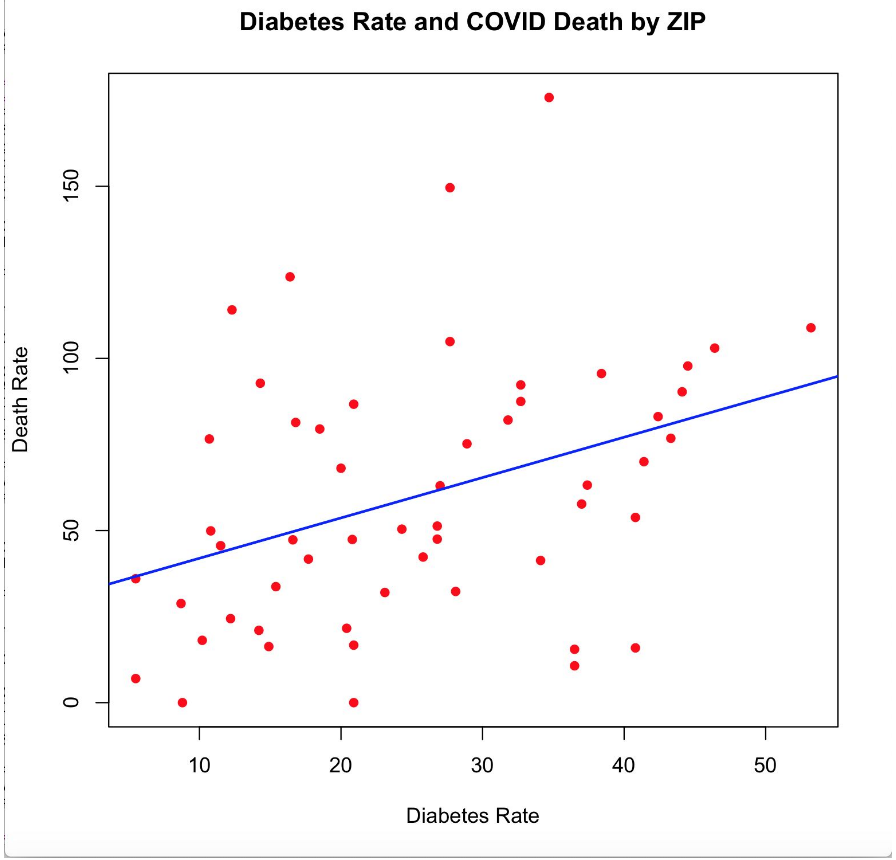
Parsimonious Model

Model:

- Looks at the rates of diabetes hospitalizations and asthma hospitalizations in each ZIP
- Statistically significant relationship with diabetes, not with asthma

Interpretation:

- An increase in the diabetes hospitalizations per 10,000 ppl leads to an increase of 1.283 in COVID deaths per 100,000 ppl
  - This indicates that diabetes is quite a serious comorbidity and is impacting community health.
- This is a statistically significant relationship with a p-value of .00534 that allows us to reject the null hypothesis.



- We can loosely infer here that diabetes has an impact on the morbidity of COVID cases in a community

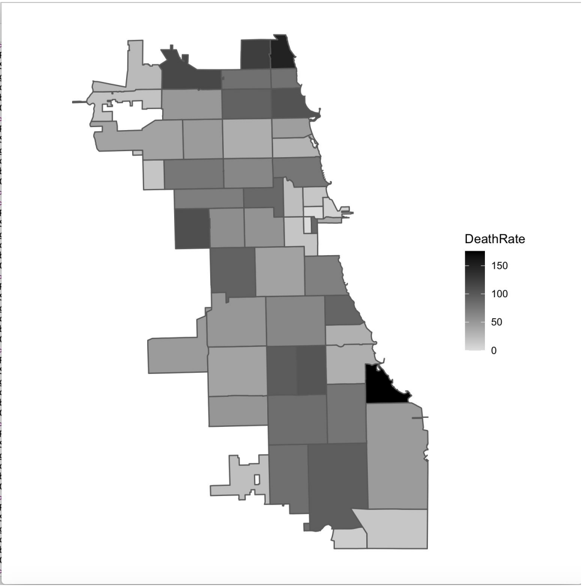
- It should be noted that Zip codes are not a perfect geographic boundary to use and some may have low populations that lead to skewed data.

- Other factors that may be related to diabetes, (i.e. Race, Gender, Lifestyle) could impact this data .

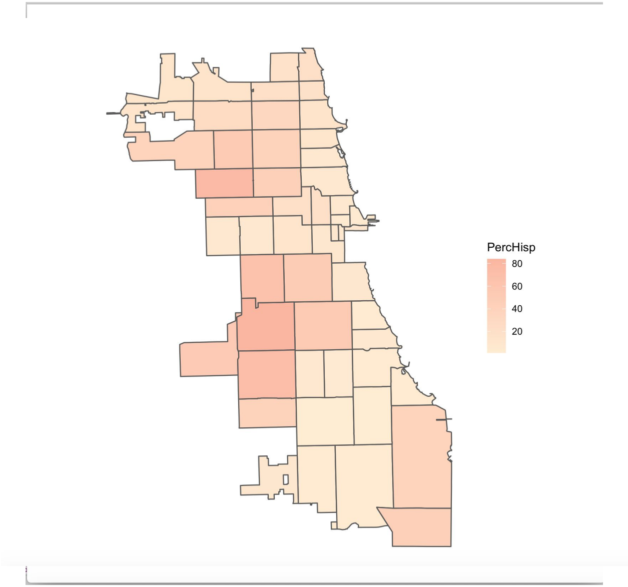
Relationship Visualized

Cook County, IL: Racial Stats by ZIP Code

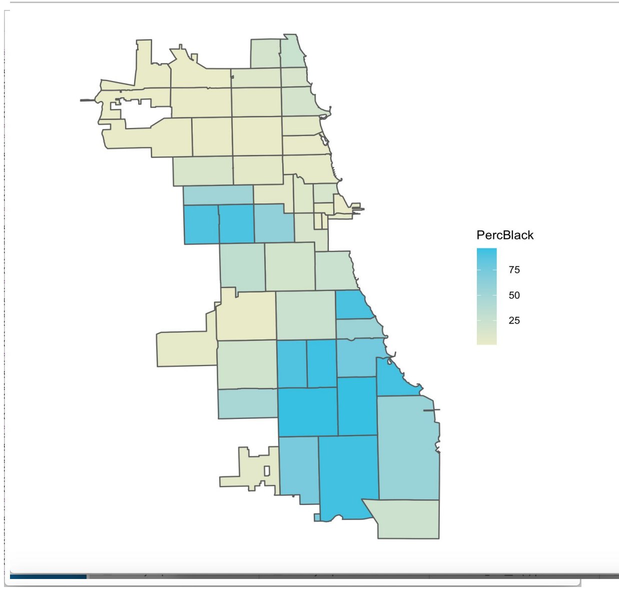
Covid 19 Deaths / 100,000 ppl



Percentage Hispanic

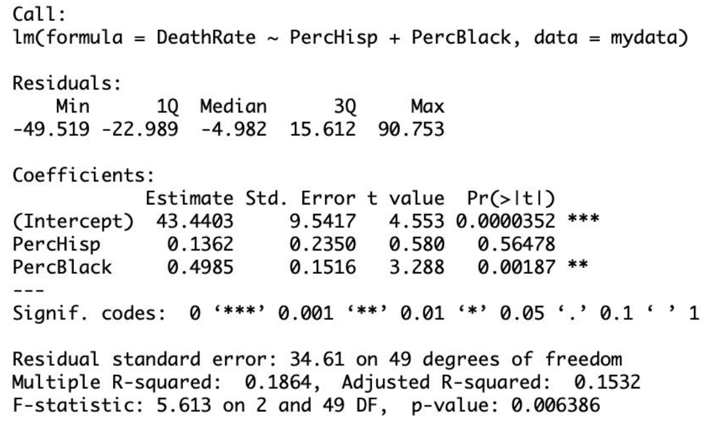


Percentage Black



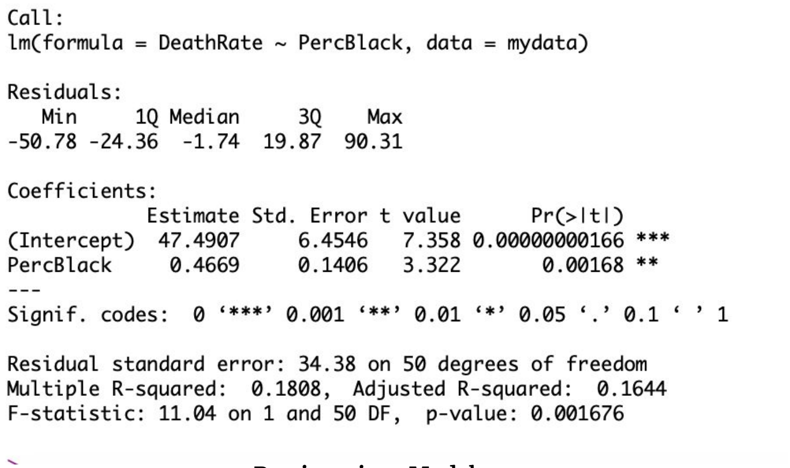
Sources: Cook County Dept. Of Health (COVID) & American Community Survey/Policy Map
http://dartmouth.policymap.com.dartmouth.idm.oclc.org/maps?p=148485&i=9868876&btd=7&period=2014-2018&cx=-87.73166187370308&cy=41.8337381642591&cz=8&iwx=41.8369&iwy=-87.6848&iwtype=place&iwtab=0

Linear Model



```
Call:
lm(formula = DeathRate ~ PercHisp + PercBlack, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-49.519 -22.989  -4.982  15.612  90.753 

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)    
(Intercept)  43.4403     9.5417   4.553 0.0000352 ***
PercHisp      0.1362     0.2350   0.580   0.56478    
PercBlack     0.4985     0.1516   3.288   0.00187 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34.61 on 49 degrees of freedom
Multiple R-squared:  0.1864,	Adjusted R-squared:  0.1532 
F-statistic: 5.613 on 2 and 49 DF,  p-value: 0.006386
```

Non-parsimonious Model



```
Call:
lm(formula = DeathRate ~ PercBlack, data = mydata)

Residuals:
   Min     1Q Median     3Q    Max 
-50.78 -24.36  -1.74  19.87  90.31 

Coefficients:
            Estimate Std. Error t value      Pr(>|t|)    
(Intercept)  47.4907     6.4546   7.358 0.00000000166 ***
PercBlack     0.4669     0.1406   3.322       0.00168 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34.38 on 50 degrees of freedom
Multiple R-squared:  0.1808,	Adjusted R-squared:  0.1644 
F-statistic: 11.04 on 1 and 50 DF,  p-value: 0.001676
```

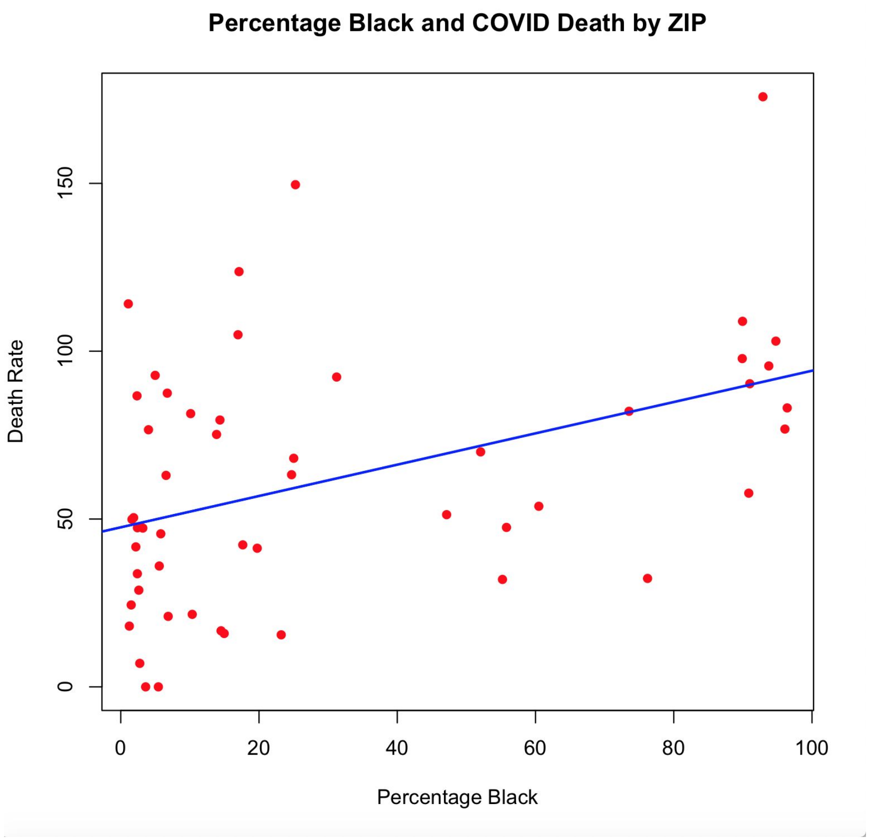
Parsimonious Model

Model:

- Looks at the percentage of the population that is Black/Hispanic in each ZIP
- Statistically significant relationship with black pop., not with Hispanic

Interpretation:

- An increase of 1% in proportion of black individuals in a community leads to an increase of .4669 in COVID deaths per 100,000 ppl
  - This indicates that there is a significant racial disparity in the effects of COVID
- This is a statistically significant relationship with a p-value of .00534



- We can infer here that COVID-19 is having a disproportionately deadly effect on the black community

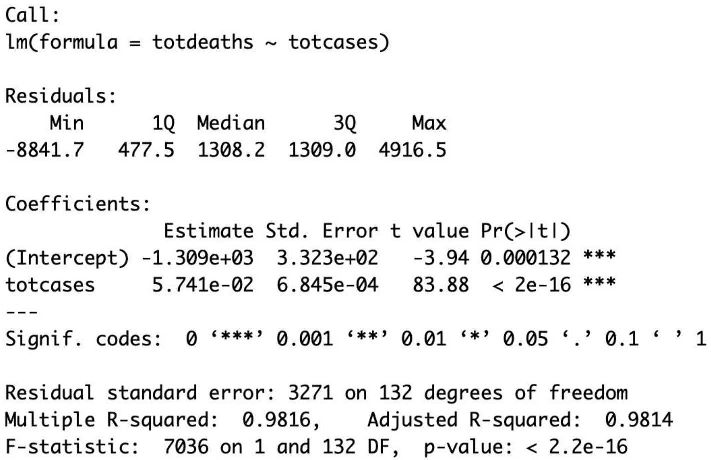
- It can be seen from this chart that almost all Zip codes with a black population above 80% have at least 50 deaths/100,000 ppl.

- This could be the result of the pattern of spreading, or other health/lifestyle factors in the community.

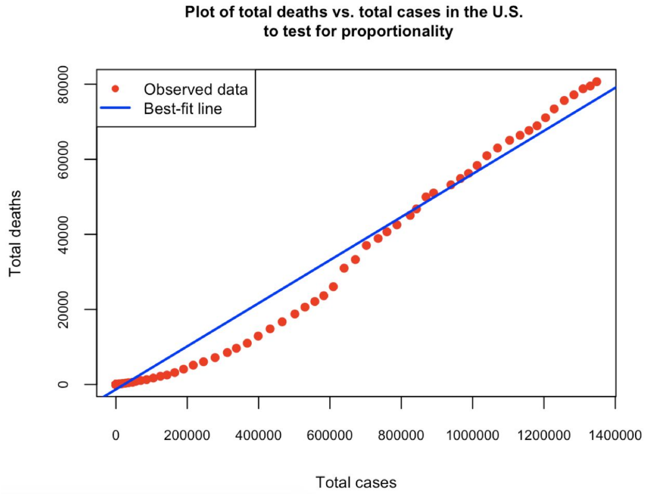
Relationship Visualized

Section 2: Broad hypotheses.

Is the number of cases proportional to the number of deaths?



```
Call:
lm(formula = totdeaths ~ totcases)

Residuals:
    Min      1Q  Median      3Q     Max 
-8841.7   477.5  1308.2  1309.0  4916.5 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.309e+03  3.323e+02   -3.94 0.000132 ***
totcases     5.741e-02  6.845e-04   83.88  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3271 on 132 degrees of freedom
Multiple R-squared:  0.9816,	Adjusted R-squared:  0.9814 
F-statistic:  7036 on 1 and 132 DF,  p-value: < 2.2e-16
```



Data source (as of 5/12/20): https://ourworldindata.org/mortality-risk-covid

Model:

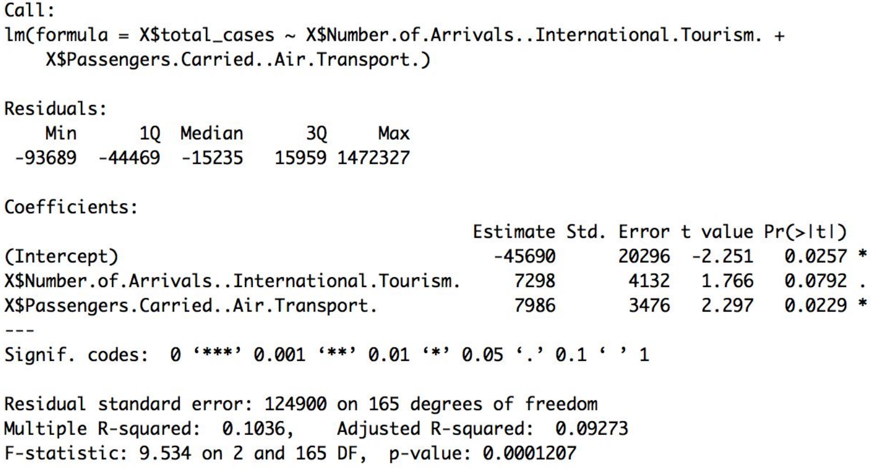
- Regress total deaths on total cases in the U.S. from January 1st through May 19th
- High-quality, statistically significant linear relationship (R-squared = 0.9816) between total deaths and total cases

Interpretation:

- An increase in case count by 1 corresponds with an increase in death count by 0.05741 in the U.S.
  - Consistent with widely known mortality rate of the disease thus far
- The number of cases is proportional to the number of deaths
- Low death count before 700K total cases, possibly because young, active people had COVID-19 but didn’t die as often as older people (as in nursing homes)

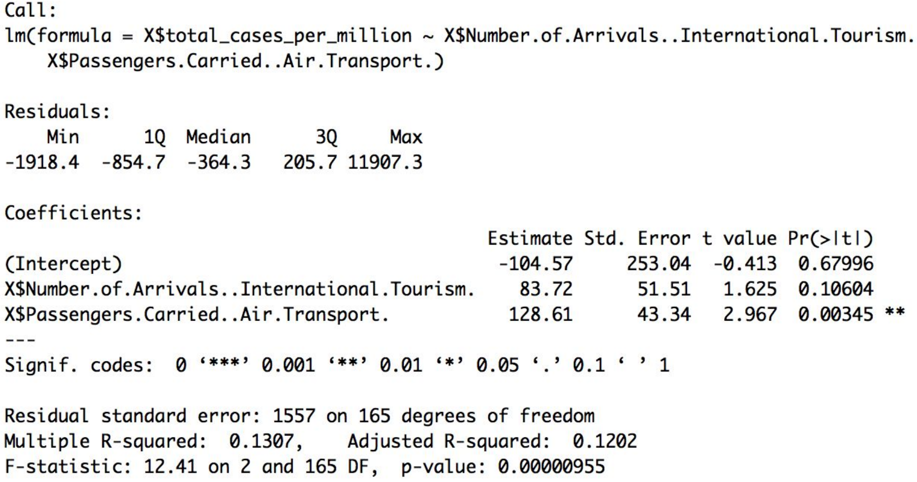
Is the number of cases proportional to the number of visitors/tourists for different countries ? (1 / 2)

Linear model for how Total Cases is explained by Number of Arrivals (International Tourism) and Passengers Carried (Air Travel)



```
Call:
lm(formula = X$total_cases ~ X$Number.of.Arrivals..International.Tourism. +
    X$Passengers.Carried..Air.Transport.)

Residuals:
    Min      1Q  Median      3Q     Max
 -93689  -44469  -15235   15959 1472327

Coefficients:
                                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                                    -45690      20296  -2.251   0.0257 *
X$Number.of.Arrivals..International.Tourism.     7298       4132   1.766   0.0792 .
X$Passengers.Carried..Air.Transport.             7986       3476   2.297   0.0229 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 124900 on 165 degrees of freedom
Multiple R-squared:  0.1036,    Adjusted R-squared:  0.09273
F-statistic: 9.534 on 2 and 165 DF,  p-value: 0.0001207
```

Linear model for how Total Cases Per Million People is explained by Number of Arrivals (International Tourism) and Passengers Carried (Air Travel)

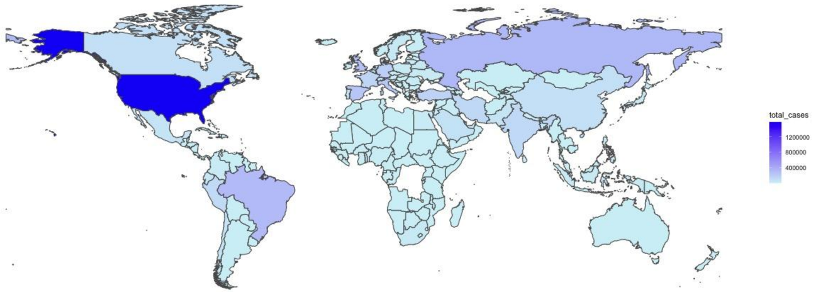


```
Call:
lm(formula = X$total_cases_per_million ~ X$Number.of.Arrivals..International.Tourism. +
    X$Passengers.Carried..Air.Transport.)

Residuals:
    Min      1Q  Median      3Q     Max
-1918.4  -854.7  -364.3   205.7 11907.3

Coefficients:
                                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                                   -104.57     253.04  -0.413  0.67996
X$Number.of.Arrivals..International.Tourism.    83.72      51.51   1.625  0.10604
X$Passengers.Carried..Air.Transport.           128.61      43.34   2.967  0.00345 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1557 on 165 degrees of freedom
Multiple R-squared:  0.1307,    Adjusted R-squared:  0.1202
F-statistic: 12.41 on 2 and 165 DF,  p-value: 0.00000955
```

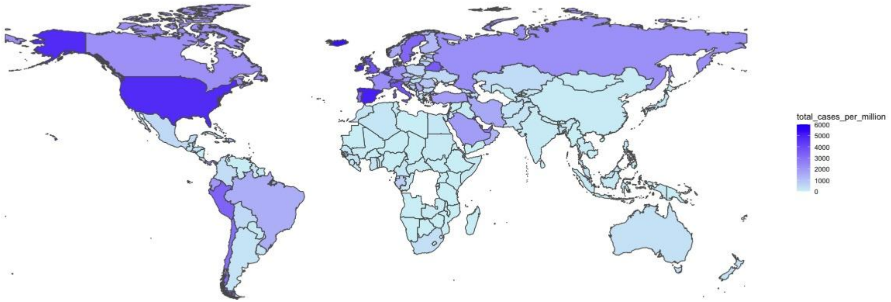
Datasets:
https://ourworldindata.org/mortality-risk-covid
https://drmkc.jrc.ec.europa.eu/inform-index/INFORM-Epidemic

Is the number of cases proportional to the number of visitors/tourists? (2 / 2)

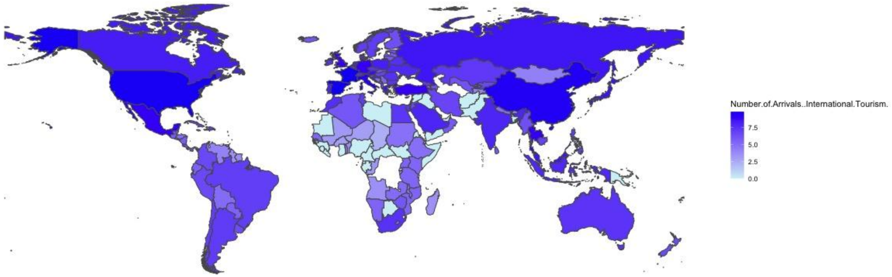
Total COVID-19 Cases by Country



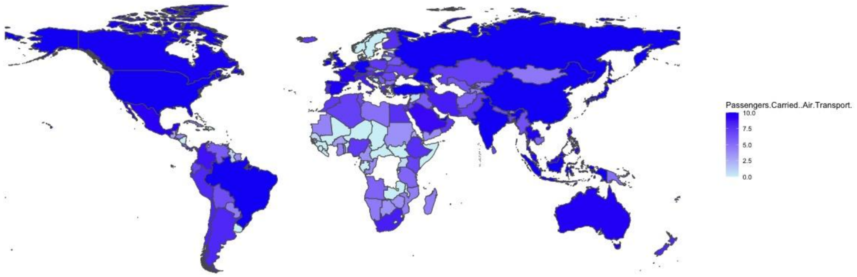
COVID-19 Cases Per Million People by Country



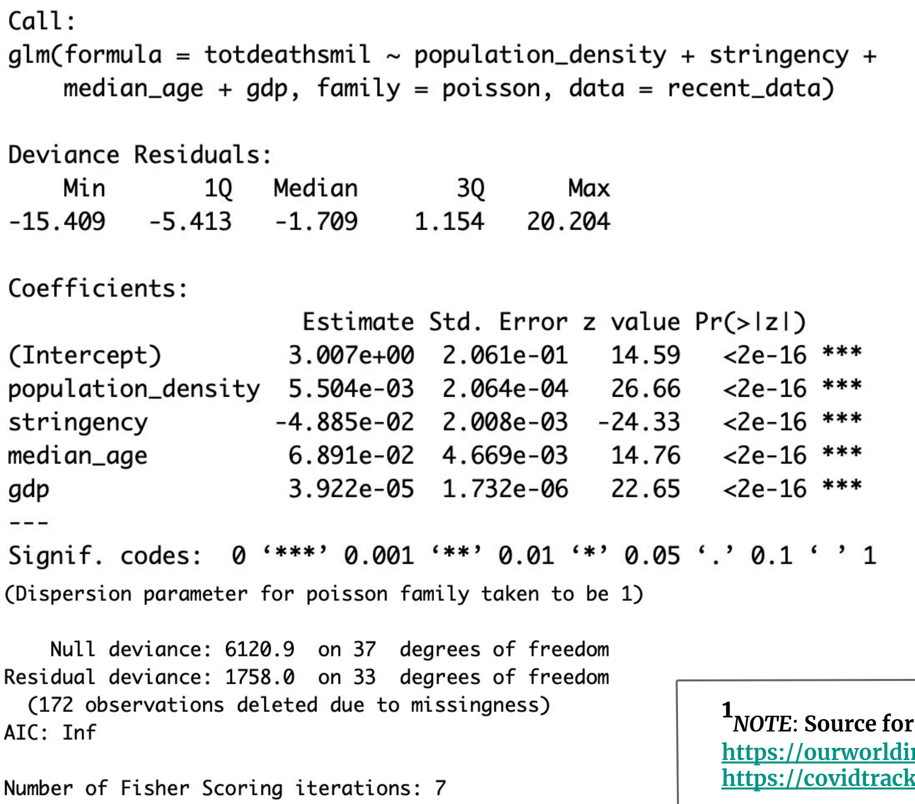
Rate of International Tourism Arrivals by Country



Rate of Air Travel Passengers by Country



How might we develop a statistically sound method to predict the expected number of deaths? (1/2)



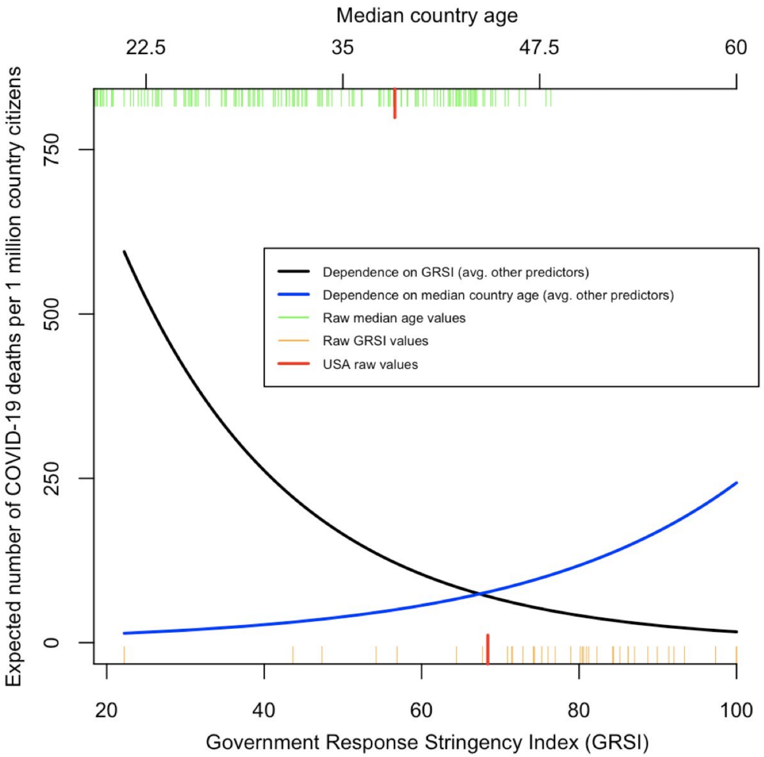
```
Call:
glm(formula = totdeathsmil ~ population_density + stringency +
    median_age + gdp, family = poisson, data = recent_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-15.409   -5.413   -1.709    1.154   20.204

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        3.007e+00  2.061e-01   14.59   <2e-16 ***
population_density 5.504e-03  2.064e-04   26.66   <2e-16 ***
stringency        -4.885e-02  2.008e-03  -24.33   <2e-16 ***
median_age         6.891e-02  4.669e-03   14.76   <2e-16 ***
gdp                3.922e-05  1.732e-06   22.65   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 6120.9  on 37  degrees of freedom
Residual deviance: 1758.0  on 33  degrees of freedom
  (172 observations deleted due to missingness)
AIC: Inf

Number of Fisher Scoring iterations: 7
```

- *A powerful tool: Poisson regression* (*glm*)
  - **Predictors:**
    - **Population density** (people per km^2)
    - **Government Response Stringency Index (GRSI)** (relative strictness of government regulations, calculated by Oxford researchers; see footnote at bottom of slide)[1]
    - **Median country age**
    - **GDP per capita**
  - **Response variable:**
    - Total deaths per 1 million country residents
  - Country data recorded as of 5/18/2020

- *Likelihood ratio test* that all coefficients are zero:
  - Subtract residual deviance from null deviance, with df difference = 3 (per R)
  - **Pseudocode**: *pchisq(null.deviance-deviance,df=3,lower.tail=F)* yields ***p*-value of 9.601615e-30**
    - **Reject the null that all coefficients are zero**

[1] *NOTE*: **Source for Government Response Stringency Index (GRSI) interpretation** at https://ourworldindata.org/grapher/covid-stringency-index, and **original Oxford GRSI source** at https://covidtracker.bsg.ox.ac.uk/; **source for original data**: https://ourworldindata.org/mortality-risk-covid.

How might we develop a statistically sound method to predict the expected number of deaths? (2/2)

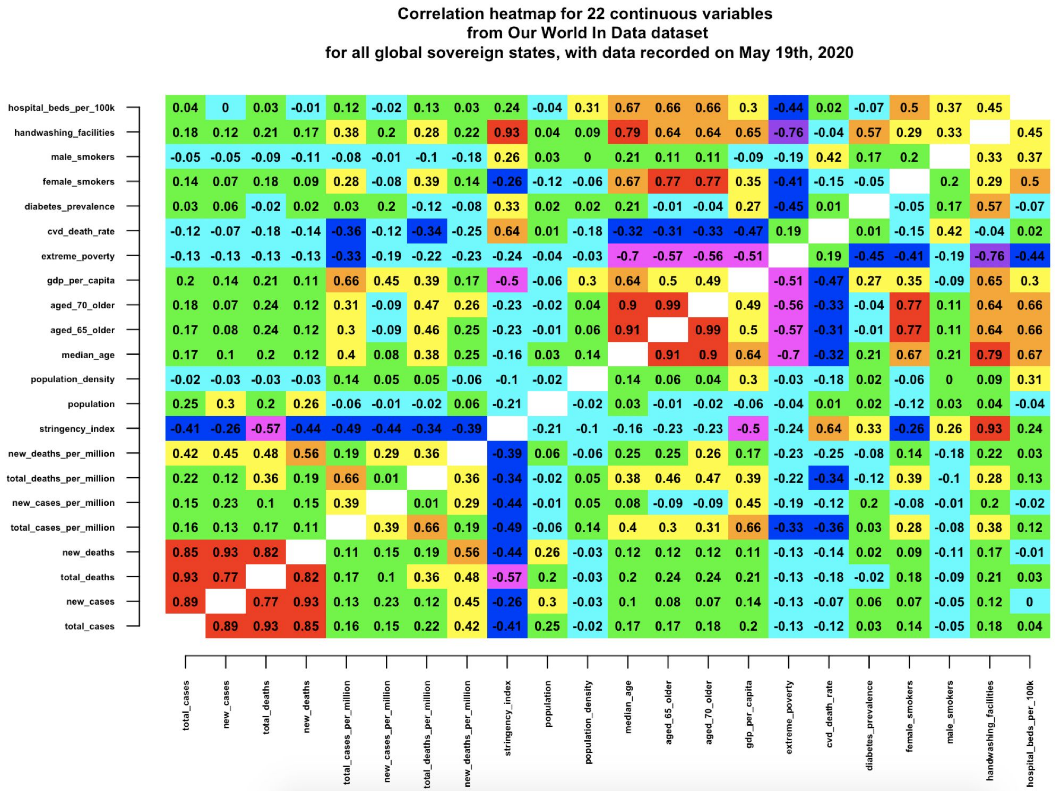


- Notable results and conclusions from plot (using *glm* and data/info sources from previous slide):
  - **Higher GRSI → lower expected COVID-19 death count per million**
    - Strict measures and restrictions working?
  - **Higher median country age → higher expected number of COVID-19 deaths**
    - Similar to linear regression model, visualized

- Results from *glm* call (previous slide):
  - **Intuitive result:** Higher population density (per previous slide) → higher death count
  - **Curious/unintuitive result:** Higher per-capita GDP → higher death count per million
    - Greater population/population density in higher-GDP countries causing this?
    - **Confounding variables at play?**

Section 3: Further predictive modelling and data exploration.

Correlation heatmap of various predictors for coronavirus death outcomes

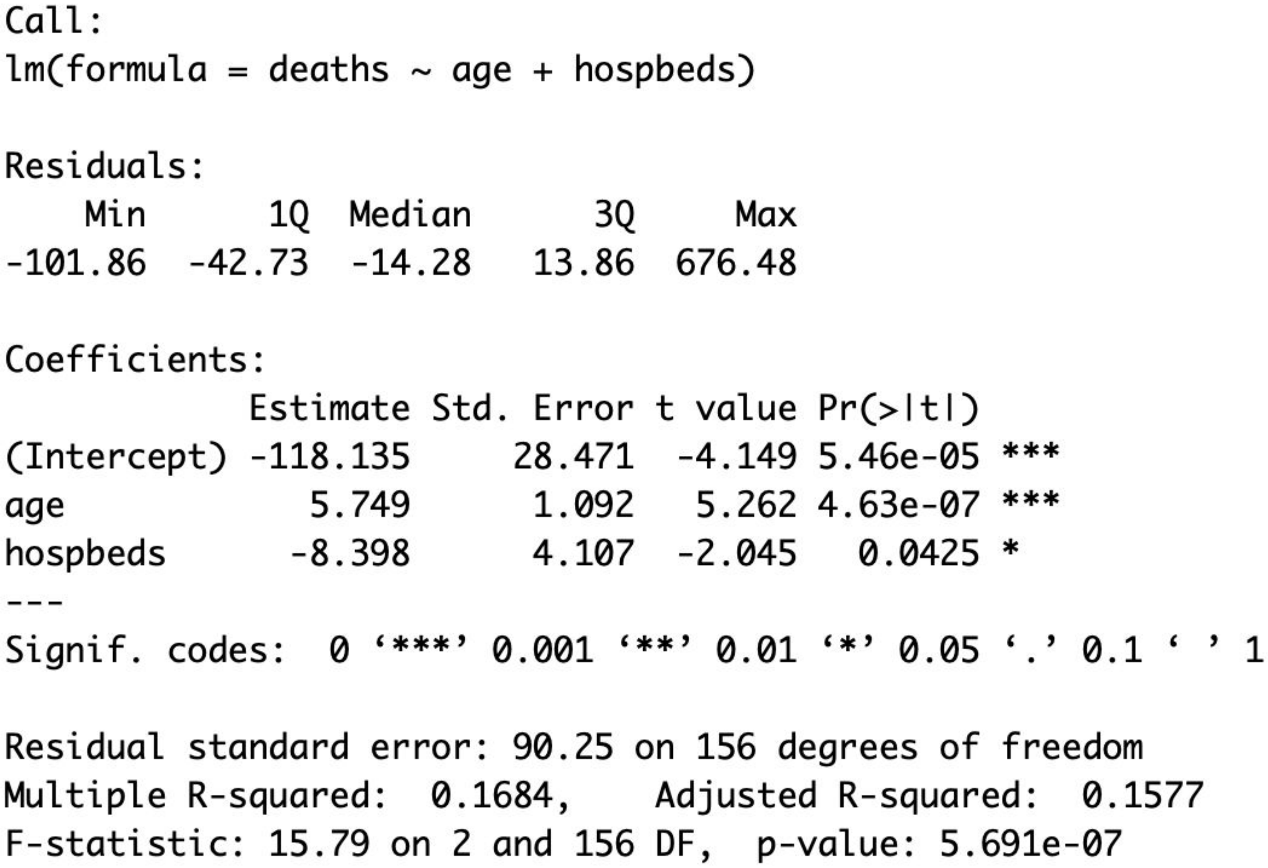


Correlation heatmap for 22 continuous variables
from Our World In Data dataset
for all global sovereign states, with data recorded on May 19th, 2020

	total_cases	new_cases	total_deaths	new_deaths	total_cases_per_million	new_cases_per_million	total_deaths_per_million	new_deaths_per_million	stringency_index	population	population_density	median_age	aged_65_older	aged_70_older	gdp_per_capita	extreme_poverty	cvd_death_rate	diabetes_prevalence	female_smokers	male_smokers	handwashing_facilities	hospital_beds_per_100k
hospital_beds_per_100k	0.04	0	0.03	-0.01	0.12	-0.02	0.13	0.03	0.24	-0.04	0.31	0.67	0.66	0.66	0.3	-0.44	0.02	-0.07	0.5	0.37	0.45	
handwashing_facilities	0.18	0.12	0.21	0.17	0.38	0.2	0.28	0.22	0.93	0.04	0.09	0.79	0.64	0.64	0.65	-0.76	-0.04	0.57	0.29	0.33		0.45
male_smokers	-0.05	-0.05	-0.09	-0.11	-0.08	-0.01	-0.1	-0.18	0.26	0.03	0	0.21	0.11	0.11	-0.09	-0.19	0.42	0.17	0.2		0.33	0.37
female_smokers	0.14	0.07	0.18	0.09	0.28	-0.08	0.39	0.14	-0.26	-0.12	-0.06	0.67	0.77	0.77	0.35	-0.41	-0.15	-0.05		0.2	0.29	0.5
diabetes_prevalence	0.03	0.06	-0.02	0.02	0.03	0.2	-0.12	-0.08	0.33	0.02	0.02	0.21	-0.01	-0.04	0.27	-0.45	0.01		-0.05	0.17	0.57	-0.07
cvd_death_rate	-0.12	-0.07	-0.18	-0.14	-0.36	-0.12	-0.34	-0.25	0.64	0.01	-0.18	-0.32	-0.31	-0.33	-0.47	0.19		0.01	-0.15	0.42	-0.04	0.02
extreme_poverty	-0.13	-0.13	-0.13	-0.13	-0.33	-0.19	-0.22	-0.23	-0.24	-0.04	-0.03	-0.7	-0.57	-0.56	-0.51		0.19	-0.45	-0.41	-0.19	-0.76	-0.44
gdp_per_capita	0.2	0.14	0.21	0.11	0.66	0.45	0.39	0.17	-0.5	-0.06	0.3	0.64	0.5	0.49		-0.51	-0.47	0.27	0.35	-0.09	0.65	0.3
aged_70_older	0.18	0.07	0.24	0.12	0.31	-0.09	0.47	0.26	-0.23	-0.02	0.04	0.9	0.99		0.49	-0.56	-0.33	-0.04	0.77	0.11	0.64	0.66
aged_65_older	0.17	0.08	0.24	0.12	0.3	-0.09	0.46	0.25	-0.23	-0.01	0.06	0.91		0.99	0.5	-0.57	-0.31	-0.01	0.77	0.11	0.64	0.66
median_age	0.17	0.1	0.2	0.12	0.4	0.08	0.38	0.25	-0.16	0.03	0.14		0.91	0.9	0.64	-0.7	-0.32	0.21	0.67	0.21	0.79	0.67
population_density	-0.02	-0.03	-0.03	-0.03	0.14	0.05	0.05	-0.06	-0.1	-0.02		0.14	0.06	0.04	0.3	-0.03	-0.18	0.02	-0.06	0	0.09	0.31
population	0.25	0.3	0.2	0.26	-0.06	-0.01	-0.02	0.06	-0.21		-0.02	0.03	-0.01	-0.02	-0.06	-0.04	0.01	0.02	-0.12	0.03	0.04	-0.04
stringency_index	-0.41	-0.26	-0.57	-0.44	-0.49	-0.44	-0.34	-0.39		-0.21	-0.1	-0.16	-0.23	-0.23	-0.5	-0.24	0.64	0.33	-0.26	0.26	0.93	0.24
new_deaths_per_million	0.42	0.45	0.48	0.56	0.19	0.29	0.36		-0.39	0.06	-0.06	0.25	0.25	0.26	0.17	-0.23	-0.25	-0.08	0.14	-0.18	0.22	0.03
total_deaths_per_million	0.22	0.12	0.36	0.19	0.66	0.01		0.36	-0.34	-0.02	0.05	0.38	0.46	0.47	0.39	-0.22	-0.34	-0.12	0.39	-0.1	0.28	0.13
new_cases_per_million	0.15	0.23	0.1	0.15	0.39		0.01	0.29	-0.44	-0.01	0.05	0.08	-0.09	-0.09	0.45	-0.19	-0.12	0.2	-0.08	-0.01	0.2	-0.02
total_cases_per_million	0.16	0.13	0.17	0.11		0.39	0.66	0.19	-0.49	-0.06	0.14	0.4	0.3	0.31	0.66	-0.33	-0.36	0.03	0.28	-0.08	0.38	0.12
new_deaths	0.85	0.93	0.82		0.11	0.15	0.19	0.56	-0.44	0.26	-0.03	0.12	0.12	0.12	0.11	-0.13	-0.14	0.02	0.09	-0.11	0.17	-0.01
total_deaths	0.93	0.77		0.82	0.17	0.1	0.36	0.48	-0.57	0.2	-0.03	0.2	0.24	0.24	0.21	-0.13	-0.18	-0.02	0.18	-0.09	0.21	0.03
new_cases	0.89		0.77	0.93	0.13	0.23	0.12	0.45	-0.26	0.3	-0.03	0.1	0.08	0.07	0.14	-0.13	-0.07	0.06	0.07	-0.05	0.12	0
total_cases		0.89	0.93	0.85	0.16	0.15	0.22	0.42	-0.41	0.25	-0.02	0.17	0.17	0.18	0.2	-0.13	-0.12	0.03	0.14	-0.05	0.18	0.04

- Notable positive correlations:
  - Total cases per million & total deaths per million (**r = 0.66**, strong linear relationship)
  - Median age & total deaths per million (**r = 0.38**)
  - Population density & total cases per million (**r = 0.14**)
  - Population density & total deaths per million (**r = 0.05**)
  - Extreme poverty & cvd_death_rate (cardiovascular disease) (**r = 0.19**)
- Notable negative correlations:
  - Handwashing facilities & extreme poverty (**r = -0.76**)
  - Proportion of pop. aged over 70 & extreme poverty (**r = -0.56**)
  - Extreme poverty & median age (**r = -0.70**)

(Data source: *Our World In Data* (updated as of May 19, 2020), https://ourworldindata.org/mortality-risk-covid)

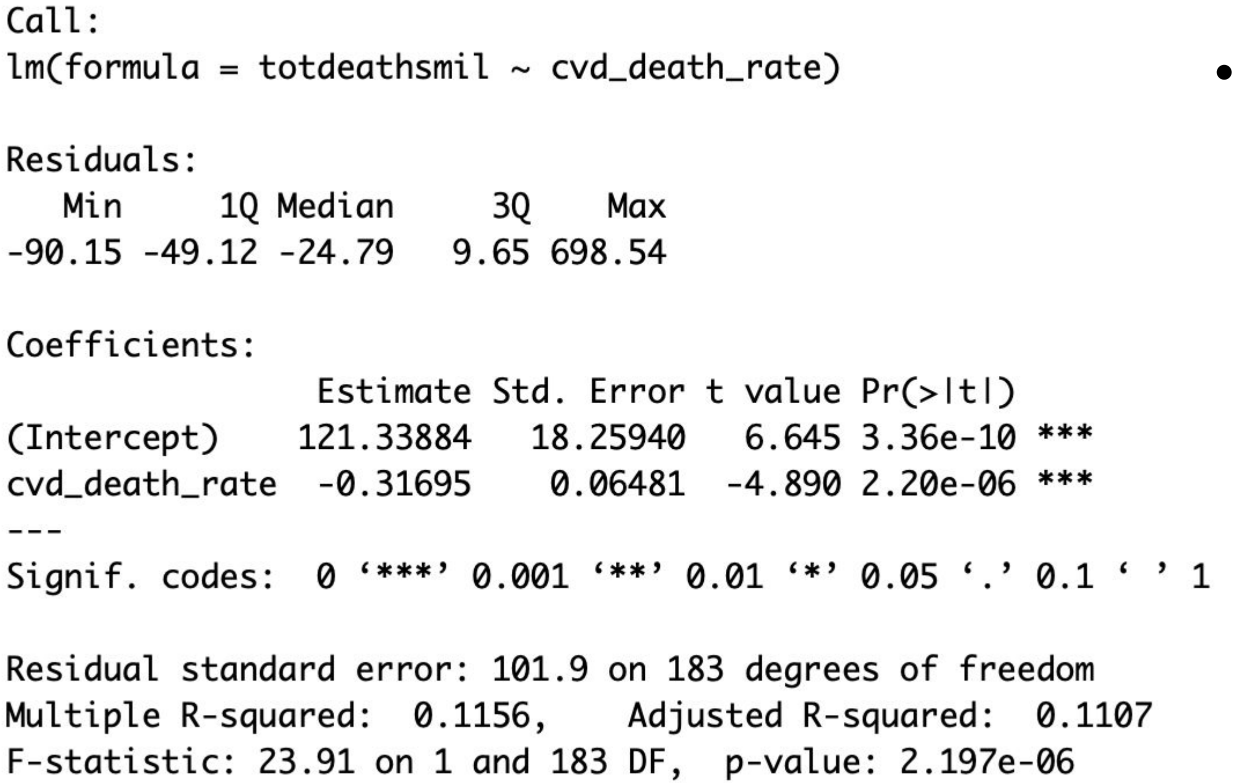


```
Call:
lm(formula = deaths ~ age + hospbeds)

Residuals:
    Min      1Q  Median      3Q     Max 
-101.86  -42.73  -14.28   13.86  676.48 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -118.135     28.471  -4.149 5.46e-05 ***
age            5.749      1.092   5.262 4.63e-07 ***
hospbeds      -8.398      4.107  -2.045   0.0425 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 90.25 on 156 degrees of freedom
Multiple R-squared:  0.1684,	Adjusted R-squared:  0.1577 
F-statistic: 15.79 on 2 and 156 DF,  p-value: 5.691e-07
```

- Parsimonious multiple linear regression model
  - ‘deaths’: deaths per 1mm citizens in country
  - ‘age’: median age in country
  - ‘hospbeds’: hospital beds per 100k citizens in country
- Interpreting ‘age’:
  - Increase in median age by 1 year correlated with 5.749 more deaths per 1mm
- Interpreting ‘hospbeds’:
  - Increase in 1 hospital bed per 100k citizens indicates 8.398 fewer deaths per 1mm
- Interpreting ‘(Intercept)’:
  - No country has a median age or number of hospital beds at 0 -> meaningless

Multiple linear regression of death count per 1 million citizens upon median age and hospital beds per 100k citizens in a given country as of May 12, 2020 (data source: *Our World In Data*, found at https://ourworldindata.org/mortality-risk-covid)

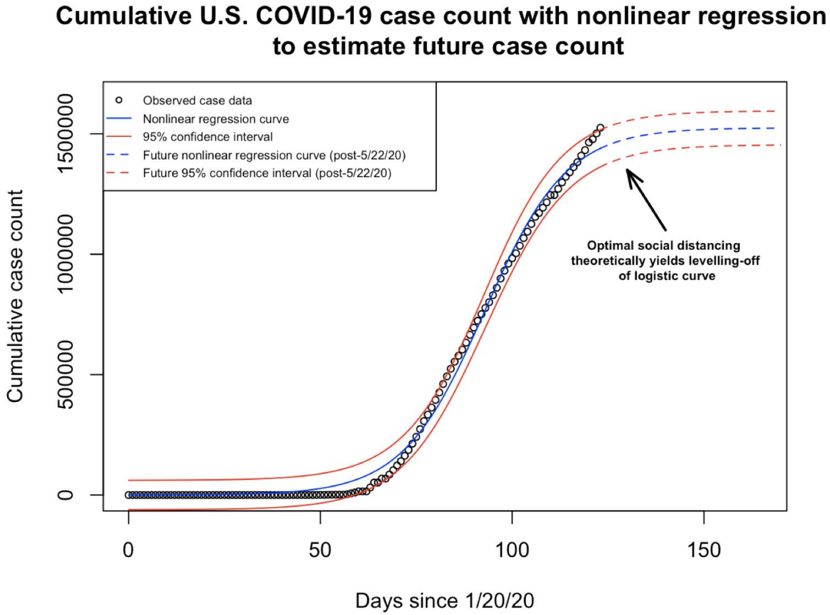
Linear regression of total COVID-19-related deaths per 1,000,000 people upon national cardiovascular disease death rate as of May 19, 2020 (data source: https://ourworldindata.org/mortality-risk-covid)



```
Call:
lm(formula = totdeathsmil ~ cvd_death_rate)

Residuals:
   Min     1Q Median     3Q    Max 
-90.15 -49.12 -24.79   9.65 698.54 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    121.33884   18.25940   6.645 3.36e-10 ***
cvd_death_rate  -0.31695    0.06481  -4.890 2.20e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101.9 on 183 degrees of freedom
Multiple R-squared:  0.1156,	Adjusted R-squared:  0.1107 
F-statistic: 23.91 on 1 and 183 DF,  p-value: 2.197e-06
```

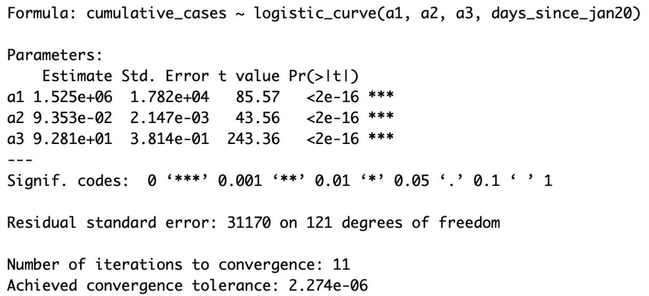
- Parsimonious model: surprising relationship
  - Increase in cardiovascular disease death rate corresponds with a decrease in total COVID-19-related deaths per million
  - The data contradict intuition
  - R-squared = 0.1156, so perhaps this relationship is not terribly concerning

A case-based nonlinear regression model that needs revamping…


Cumulative U.S. COVID-19 case count with nonlinear regression to estimate future case count

Optimal social distancing theoretically yields levelling-off of logistic curve

- Logistic equation for *nls* model (see sources on bottom right): $$c = \frac{a_1}{1 + e^{-a_2(t-a_3)}}$$
- Clearly, case *estimates* are too high before day 70 and significantly too low by around day 110
  - **Here, the logistic equation overfits and does <u>not</u> provide useful prediction capability, despite its frequent use in infectious disease modeling (as seen in sources on bottom right)**
- We must select a better nonlinear regression equation (in the next model) than what is shown below:

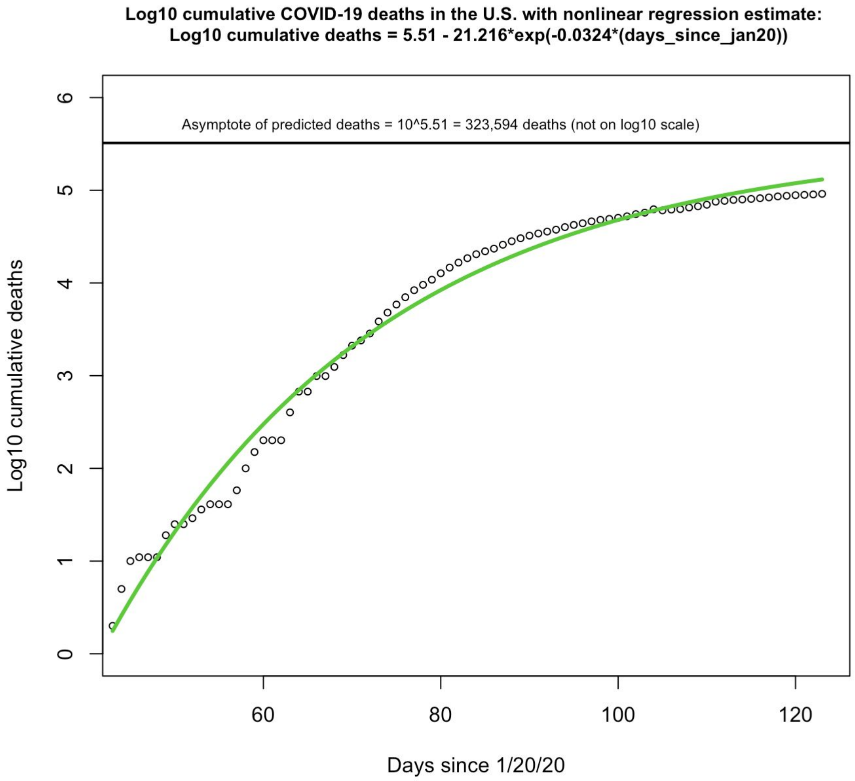


```
Formula: cumulative_cases ~ logistic_curve(a1, a2, a3, days_since_jan20)

Parameters:
    Estimate Std. Error t value Pr(>|t|)
a1 1.525e+06  1.782e+04   85.57   <2e-16 ***
a2 9.353e-02  2.147e-03   43.56   <2e-16 ***
a3 9.281e+01  3.814e-01  243.36   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31170 on 121 degrees of freedom

Number of iterations to convergence: 11
Achieved convergence tolerance: 2.274e-06
```

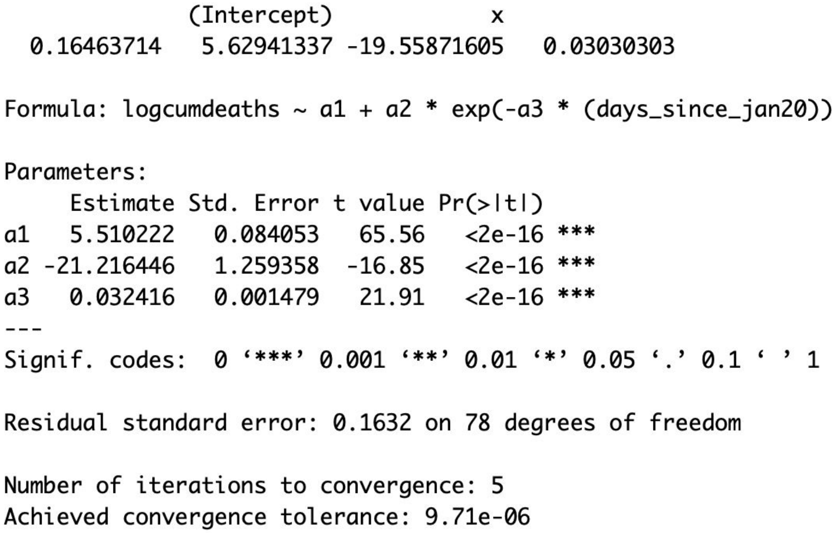
Data source:
https://data.humdata.org/dataset/coronavirus-covid-19-cases-and-deaths

Source for info on logistic function ("Logistic growth" section):
https://xaktly.com/LogisticFunctions.html
Further source for info on logistic function's use in modeling infection spread: http://www.nlreg.com/aids.htm

A proper nonlinear regression approach to model and predict the number of COVID-19 deaths in the U.S.



Log10 cumulative COVID-19 deaths in the U.S. with nonlinear regression estimate:
Log10 cumulative deaths = 5.51 - 21.216*exp(-0.0324*(days_since_jan20))

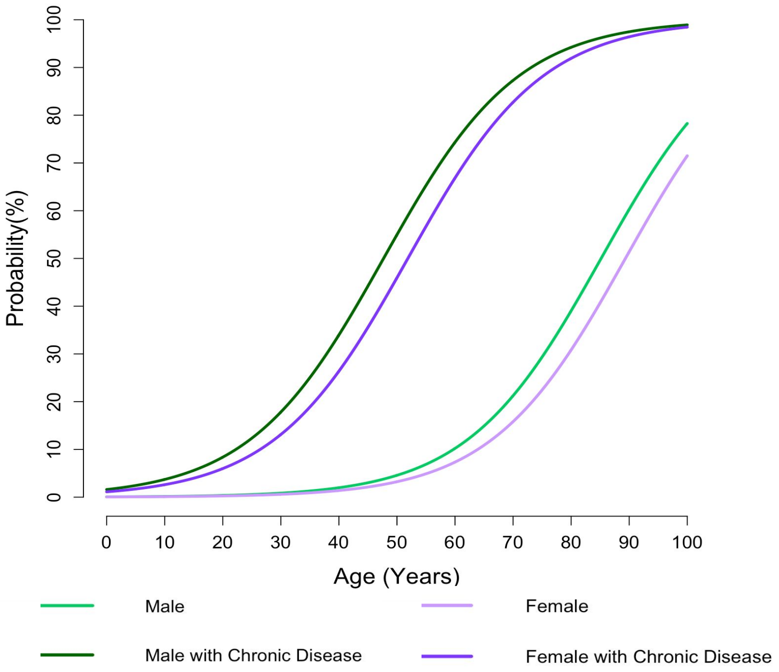


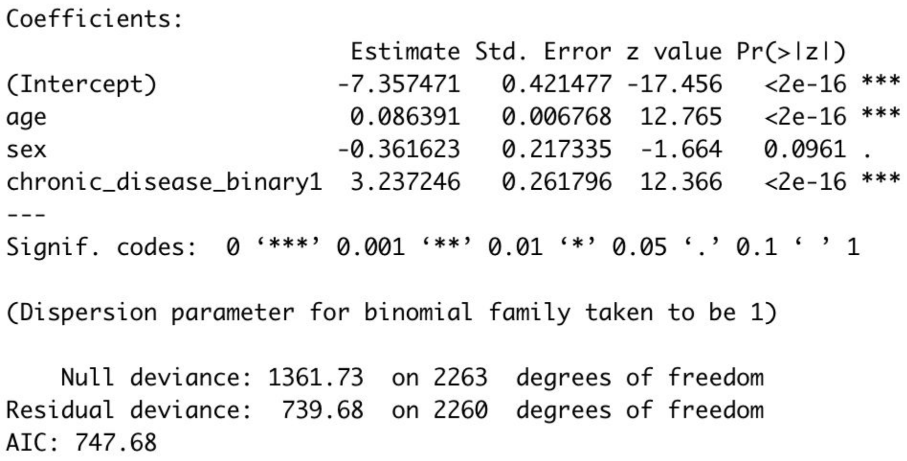
```
                 (Intercept)             x
 0.16463714   5.62941337 -19.55871605   0.03030303

Formula: logcumdeaths ~ a1 + a2 * exp(-a3 * (days_since_jan20))

Parameters:
     Estimate Std. Error t value Pr(>|t|)
a1   5.510222   0.084053   65.56   <2e-16 ***
a2 -21.216446   1.259358  -16.85   <2e-16 ***
a3   0.032416   0.001479   21.91   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1632 on 78 degrees of freedom

Number of iterations to convergence: 5
Achieved convergence tolerance: 9.71e-06
```

- Data source: https://data.humdata.org/dataset/coronavirus-covid-19-cases-and-deaths
- Asymptote is equivalent to coefficient a1: taking the limit as t → ∞, we get asymptotically 10^5.51 = 323,594 deaths in the long run
  - To what extent can we trust this result?
- We do not have an interpretation before about 40 days following 1/20/2020—death counts were at 0, so the log is not defined
- Deaths are clearly flattening out rate-wise since the beginning of the pandemic in the U.S.

Utilizing patient level data to find odds ratios, logistic regression, and likelihood ratio tests to predict individual outcomes (1/3)

Probability of COVID-19 Case Resulting in Death as a Function of Age





```
Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)            -7.357471   0.421477 -17.456   <2e-16 ***
age                     0.086391   0.006768  12.765   <2e-16 ***
sex                    -0.361623   0.217335  -1.664   0.0961 .
chronic_disease_binary1  3.237246   0.261796  12.366   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1361.73  on 2263  degrees of freedom
Residual deviance:  739.68  on 2260  degrees of freedom
AIC: 747.68
```

Xu, B., Gutierrez, B., Mekaru, S. *et al.* Epidemiological data from the COVID-19 outbreak, real-time case information. *Sci Data* 7, 106 (2020). https://doi.org/10.1038/s41597-020-0448-0

Utilizing patient level data to find odds ratios, logistic regression, and likelihood ratio tests to predict individual outcomes (2/3)

- OR for age: exp(0.086391)=1.090233
  - For a one year increase in patient age we expect to see a 10.7% increase in the odds of the outcome being death

- OR for sex: exp(-0.361623)=0.6965449
  - A female patient outcome resulting in death has 37% lower odds than in males

- OR for chronic disease: exp(3.237246)=25.4635
  - A patient with chronic disease has 25.5 times the odds of dying than one without a chronic disease

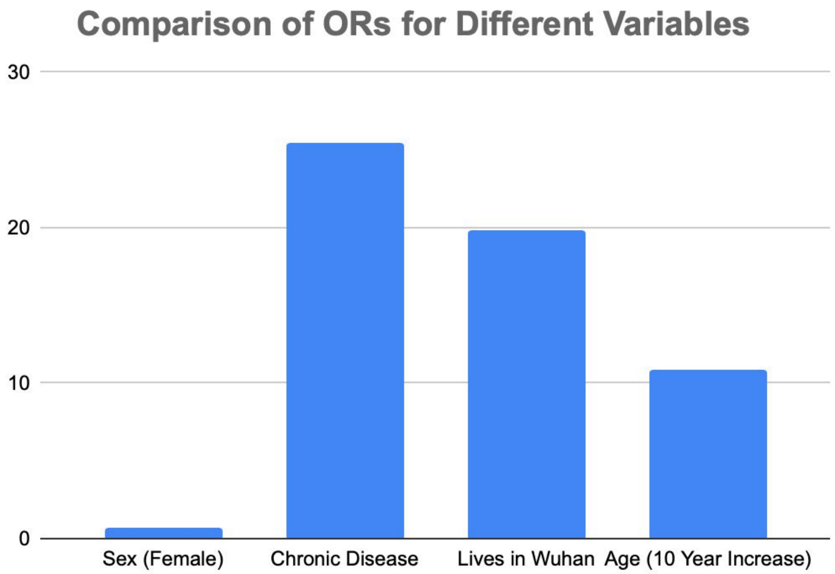
- Likelihood Ratio Test: testing null that all coefficients are 0
  - P-value: 1.671506e-134

Xu, B., Gutierrez, B., Mekaru, S. *et al.* Epidemiological data from the COVID-19 outbreak, real-time case information. *Sci Data* 7, 106 (2020). https://doi.org/10.1038/s41597-020-0448-0

Utilizing patient level data to find odds ratios, logistic regression, and likelihood ratio tests to predict individual outcomes (3/3)

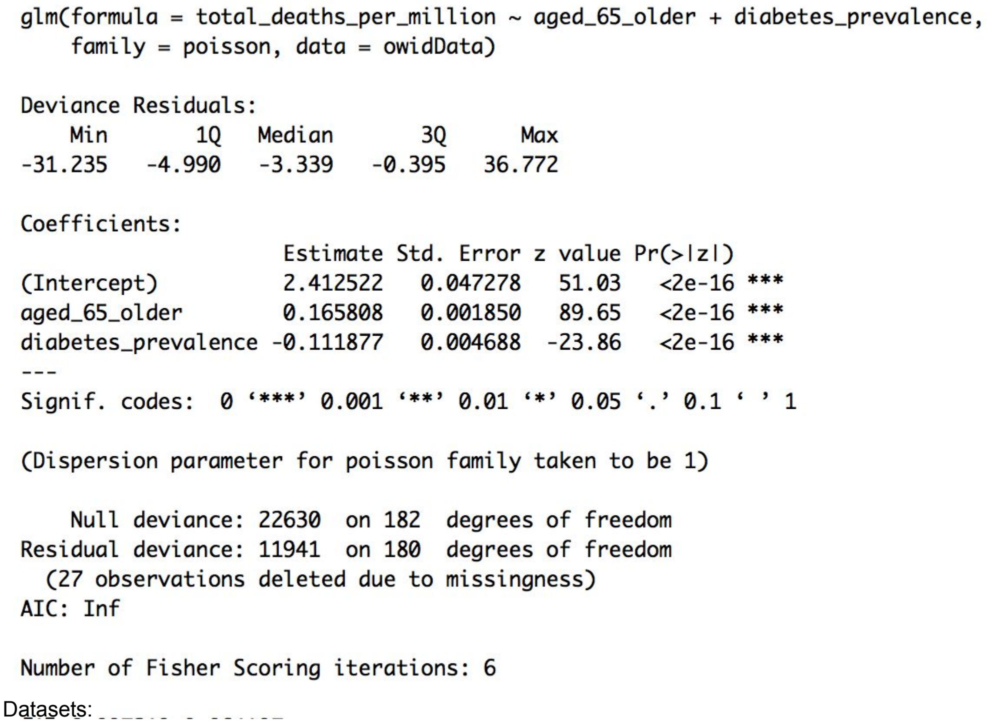
Similar outcome (death vs recovered) logistic regression run with independent variable whether the patient lives in Wuhan, China. Much smaller sample size, but significant results:

- Odds Ratio: exp(2.9874)=19.83405
  - A patient living in Wuhan has 19.8 times the odds of dying than one living somewhere else
- LR test: pchisq(190.78-149.22, df=1, lower.tail = F)
  - P-value: 1.143083e-10


Comparison of ORs for Different Variables

Xu, B., Gutierrez, B., Mekaru, S. *et al.* Epidemiological data from the COVID-19 outbreak, real-time case information. *Sci Data* 7, 106 (2020). https://doi.org/10.1038/s41597-020-0448-0

How do diabetes prevalence and old age affect the rate of death counts?



```
glm(formula = total_deaths_per_million ~ aged_65_older + diabetes_prevalence,
    family = poisson, data = owidData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-31.235   -4.990   -3.339   -0.395   36.772

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          2.412522   0.047278   51.03   <2e-16 ***
aged_65_older        0.165808   0.001850   89.65   <2e-16 ***
diabetes_prevalence -0.111877   0.004688  -23.86   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 22630  on 182  degrees of freedom
Residual deviance: 11941  on 180  degrees of freedom
  (27 observations deleted due to missingness)
AIC: Inf

Number of Fisher Scoring iterations: 6
```

Interpretation of Coefficients:

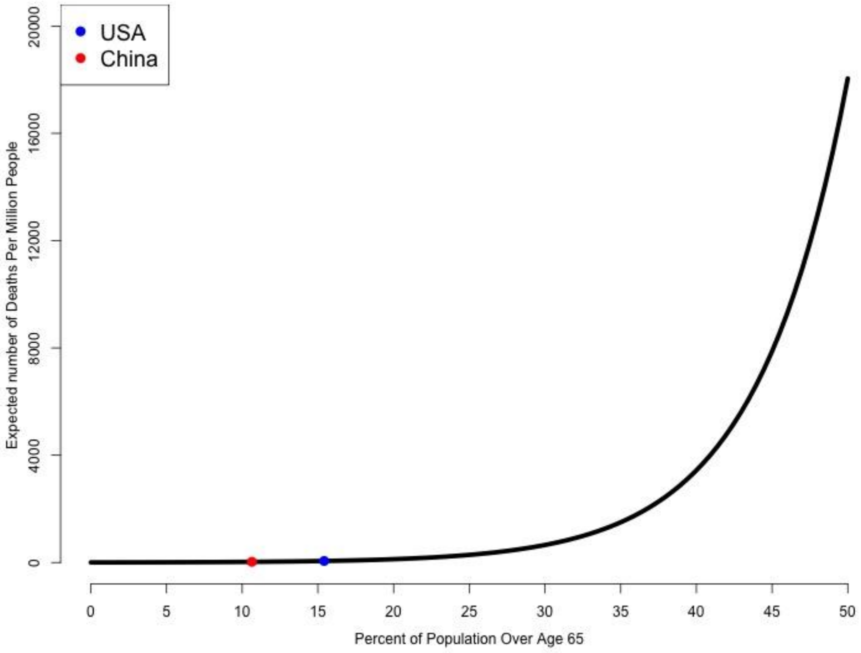
- For every percent increase in *the percentage of the population older than 65*, the *total deaths per million* is expected to increase by e^(0.166) (approx. 1.18)
- For every percent increase in *the percentage of the population with diabetes*, the *total deaths per million* is expected to increase by e^(-0.112) (approx. 0.89)

Interpretation of p-values:

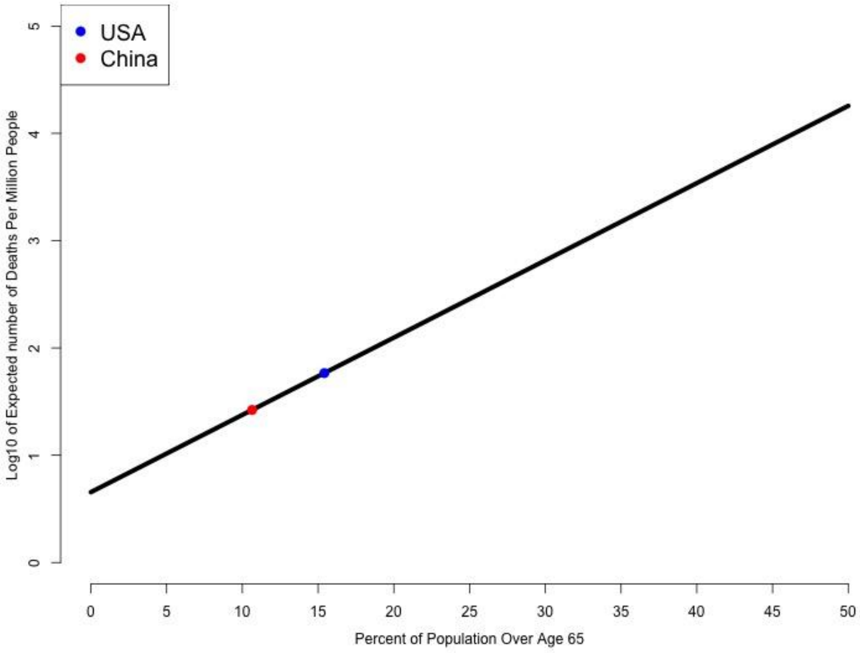
- Under the null hypothesis that each coefficient is equal to zero, we would expect to see coefficients at or more extreme as the ones we observed with a probability of approx. 2e-16
- Therefore both coefficients are statistically significant and contribute to the model

Datasets:
https://ourworldindata.org/mortality-risk-covid
https://drmkc.jrc.ec.europa.eu/inform-index/INFORM-Epidemic

Expected Deaths as a Function of the Percentage of Population Older than 65 Years

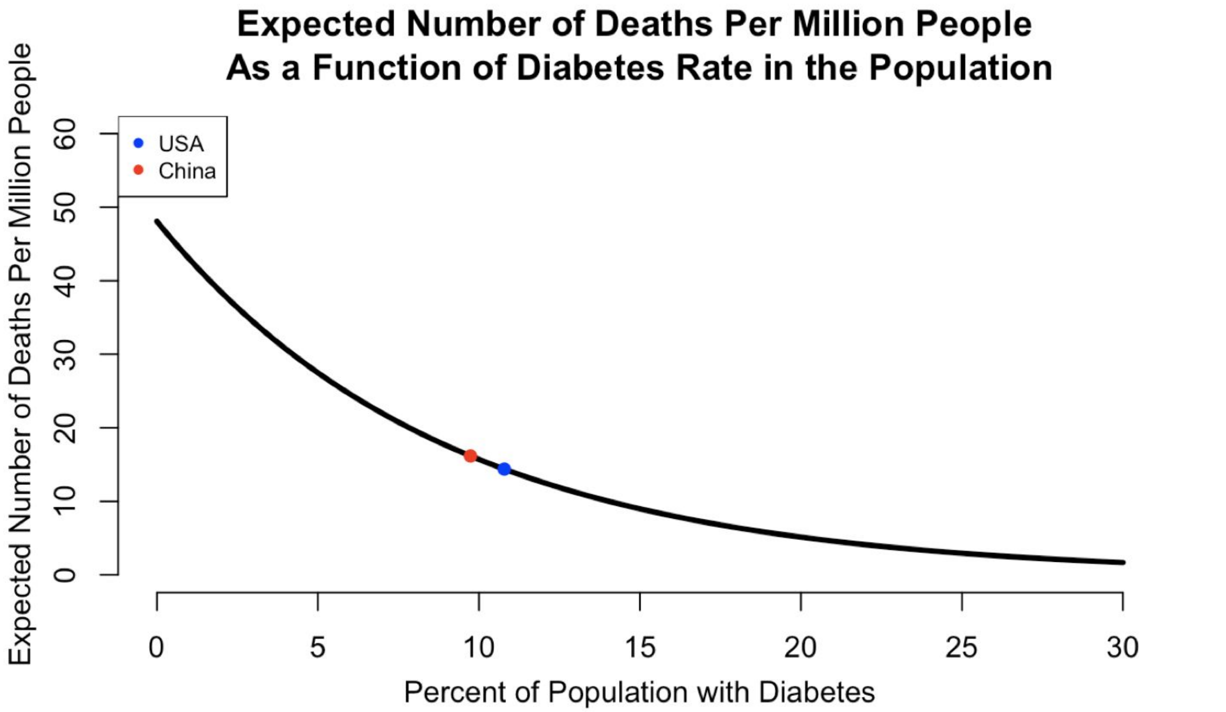


Expected Number of Deaths Per Million People
As a Function of Percent of Population Older than 65

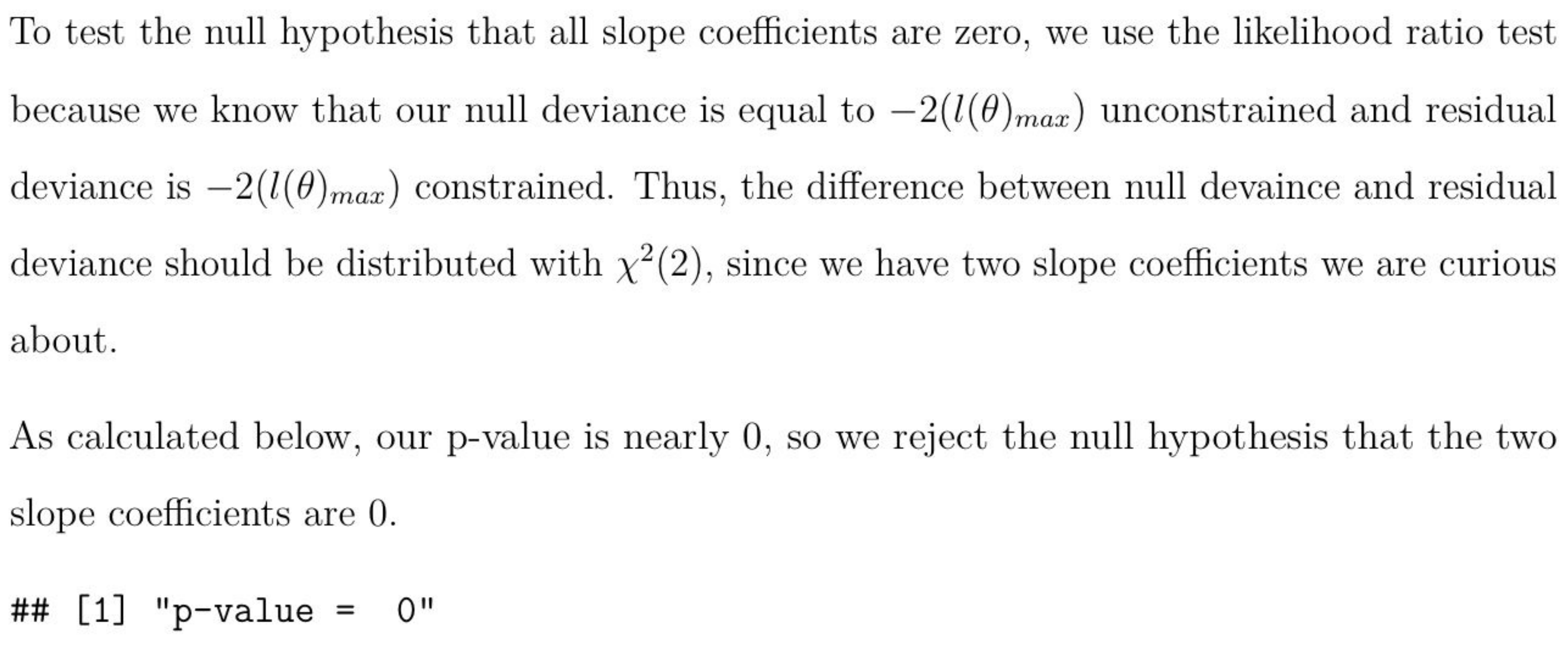


Log of Expected Number of Deaths Per Million People
As a Function of Percent of Population Older than 65

Expected Deaths as a Function of the Percentage of Population with Diabetes



Testing if the GLM Coefficients are zero (Likelihood Ratio Test)



To test the null hypothesis that all slope coefficients are zero, we use the likelihood ratio test because we know that our null deviance is equal to $-2(l(\theta)_{max})$ unconstrained and residual deviance is $-2(l(\theta)_{max})$ constrained. Thus, the difference between null devaince and residual deviance should be distributed with $\chi^2(2)$, since we have two slope coefficients we are curious about.

As calculated below, our p-value is nearly 0, so we reject the null hypothesis that the two slope coefficients are 0.

```
## [1] "p-value =  0"
```

Section 4: Conclusions.

What are the key takeaways from our data exploration and statistical analysis?

- Testing is still lacking in many parts of the world—the “case count” is not accurate
- The data are messy (with many missing values and inconclusive column names) and updated daily
- Dartmouth students like us are *not* epidemiologists
  - However, we *can* become more informed by performing and understanding these analyses while acknowledging our inexperience
- Organizations like the CDC mix probable and true deaths together...
  - Combined with under-diagnosis issues, can we trust open-source (or *any*) COVID-19 data?
- Cases will plateau if social distancing and strict governmental policy continue
  - Is this likely?
- Poorer countries with higher rates of chronic disease (like CVD) have been hit hardest (in terms of death rate)
- Big question: How might we make the data more *reliable*?
  - Increasing testing capacity (to make case data more accurate) is one option

Thank you!

We hope you found our analysis and takeaways useful in this uncertain, trying time.