

## Work for the rest of the semester (4/20 to 5/5)

### Assignment #5

Due in Sakai 3pm EST on 4/28. (Solutions will be posted on Sakai once everyone turns in the homework)

### The final exam (25 pts+5 pts toward the final grade):

1. in-class exam on chap 6 (12:30 to 1:45pm EST on 4/29): 15 pts
2. covid-19 project: (10 pts+5 bonus points); due 10am EST on 5/5. Below is the state assignment (columns are State, Student's last name, # of counties, total # of counties)

Illinois	Alutto	102	174	Arizona	Milikich	15	345
Wisconsin	Alutto	72		New Mexico	Milikich	33	
Maine	Corrigan	16	148	Texas	Milikich	254	
Pennsylvania	Corrigan	67		Maryland	Mudd	24	91
Vermont	Corrigan	14		Pennsylvania	Mudd	67	
Alabama	Cu	67	226	Alaska	Mummery	19	82
Georgia	Cu	159		California	Mummery	58	
Delaware	Davin	3	86	Hawaii	Mummery	5	81
New Jersey	Davin	21		Massachusetts	O'Connell	14	81
New York	Davin	62		New York	O'Connell	62	
Florida	Demirci	67	182	Rhode Island	O'Connell	5	
Missouri	Demirci	115		Montana	Piland	56	102
North Carolina	Farrelly	100	146	Nevada	Piland	17	
South Carolina	Farrelly	46		Utah	Piland	29	
Indiana	Gillen	92	197	Louisiana	Santiaguel	64	141
Kansas	Gillen	105		Oklahoma	Santiaguel	77	
North Dakota	Harding Bradley	53	119	Hawaii	Tang	5	80
South Dakota	Harding Bradley	66		Oregon	Tang	36	
Virginia	Hatcher	133	188	Washington	Tang	39	
West Virginia	Hatcher	55		Arkansas	Ulrich	75	157
Iowa	Huang	99	192	Mississippi	Ulrich	82	
Nebraska	Huang	93		Colorado	Yan	64	131
Michigan	Liang	83	258	Idaho	Yan	44	
Minnesota	Liang	87		Wyoming	Yan	23	
Ohio	Liang	88		Connecticut	Zhao	8	80
Kentucky	Lipa	120	215	New Hampshire	Zhao	10	
Tennessee	Lipa	95		New York	Zhao	62	

### Instructions on the final project


#### Plan for at least 8 hrs to work on the project

- ~ 4 hrs on data preprocessing and merging
- ~ 2 hrs on running the analysis
- ~ 2 hrs on reporting and submission

### Data source:

1. covid19.txt: covid-19 total number of cases,
  - subset the data for the states assigned to you
  - include all dates and run the following codes to get the days from 1/1/2020 (the highlighted yellow) areas are placeholders that will need to be customized).

```
mydate <- factor(coviddata$date)
mydate<- as.Date(mydate, format = "%Y/%m/%d")
coviddata$date<- mydate- as.Date("2020/1/1", "%Y/%m/%d")
```
2. PovertyEstimates.xls
3. Unemployment.xls
4. Education.xls
5. PopulationEstimates.xls
6. Demographic info (age, gender and race)
  - Go to [https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html#par\\_textimage\\_1383669527](https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html#par_textimage_1383669527)
  - Scroll down to Annual County Resident Population Estimates by Age, Sex, Race, and Hispanic Origin: April 1, 2010 to July 1, 2018 (CC-EST2018-YOURCOMBINEDDATA) (see a screenshot below).  
Annual County Resident Population Estimates by Age, Sex, Race, and Hispanic Origin: April 1, 2010 to July 1, 2018 (CC-EST2018-ALLDATA)

 File Layout [<1.0 MB]

Choose a State to View.

United States	Kansas	North Carolina
Alabama	Kentucky	North Dakota
Alaska	Louisiana	Ohio
Arizona	Maine	Oklahoma
Arkansas	Maryland	Oregon
California	Massachusetts	Pennsylvania
Colorado	Michigan	Rhode Island
Connecticut	Minnesota	South Carolina
Delaware	Mississippi	South Dakota
District of Columbia	Missouri	Tennessee
Florida	Montana	Texas
Georgia	Nebraska	Utah
Hawaii	Nevada	Vermont
Idaho	New Hampshire	Virginia
Illinois	New Jersey	Washington
Indiana	New Mexico	West Virginia
Iowa	New York	Wisconsin
		Wyoming

- Select the states assigned to download the files.
- Run the following R codes to obtain the demographic data (the highlighted areas are placeholders that will need to be customized).

```

demo0<-read.csv("youfilename.csv",header=T)
used.col <-
c('STNAME','CTYNAME','AGEGRP','TOT_POP','TOT_MALE','TOT_FEMALE',
'WA_MALE','WA_FEMALE','BA_MALE','BA_FEMALE','AA_MALE','AA_FEMALE',
'H_MALE','H_FEMALE')
demo<- demo0[demo0$YEAR==11,]
demo<- demo[, used.col]

total <- demo[demo$AGEGRP==0,];
Pmale<- total$TOT_MALE/total$TOT_POP

Pwhite<- (total$WA_MALE+total$WA_FEMALE)/total$TOT_POP
Pblack<- (total$BA_MALE+total$BA_FEMALE)/total$TOT_POP
Pasian<- (total$AA_MALE+total$AA_FEMALE)/total$TOT_POP
Phispanic<- (total$H_MALE+total$H_FEMALE)/total$TOT_POP

age<-matrix(demo[, 4], ncol=19,byrow=T)
Page<- as.data.frame(age[,-1]/age[,1]);
colnames(Page)= c(paste0("Page", 1:18))
demo.final<-cbind(total[, c(1:2,4)], Pmale, Pwhite, Pblack,
Pasian, Phispanic, Page);
write.csv(demodata,'DemoData.csv',row.names = F)

```

7. Weather data: county-level maximum and minimum, temperature, and precipitation data
  - Go to <https://www.ncdc.noaa.gov/cag/county/mapping/1/tmax/202003/1/value>
  - Select the states assigned to you one by one, select year 2020, Month “March”, time scale “1-month”, “maximum temperature” (see below for a screenshot)

## County Mapping

Choose from the options below and click "Plot" to create a map. Select [Temperature](#) and [Precipitation Maps](#) are available for download.

State:

Parameter:

Year:



Month:



Time Scale:

*Please note, Degree Days and Palmer not available for Counties. These data available for [bulk download](#).*

Plot

- The downloadable data are below the plot (see below for a screenshot)

Download Table Data: [XML](#)  

Download All Months/Years: [XML](#)  

COUNTY	VALUE	RANK (126 YEARS)	1901-2000 MEAN	ANOMALY
Autauga County, AL	76.9°F	125	68.3°F	8.6°F
Baldwin County, AL	78.2°F	124	70.5°F	7.7°F
Barbour County, AL	75.8°F	120	69.1°F	6.7°F
Bibb County, AL	73.3°F	116	66.9°F	6.4°F
Blount County, AL	69.7°F	111	64.4°F	5.3°F
Bullock County, AL	75.1°F	120	68.3°F	6.8°F
Butler County, AL	77.9°F	124	69.5°F	8.4°F
Calhoun County, AL	70.3°F	112	65.0°F	5.3°F
Chambers County, AL	71.6°F	114	66.3°F	5.3°F
Cherokee County, AL	68.1°F	106	64.0°F	4.1°F
Chilton County, AL	74.9°F	121	67.0°F	7.9°F
Choctaw County, AL	76.7°F	120	69.6°F	7.1°F
Clarke County, AL	77.2°F	120	70.0°F	7.2°F

- Repeat the above step to download the minimum temperature and precipitation data.
- In each of the download the file, keep the state, county, value columns only, merge the three files into one file 'weather.csv' with state, county, maximum temperature, minimum temperature, and precipitation values.

### Merging data by state and county

Merge the 7 data sources (i.e., covid19.txt, PovertyEstimates.xls, Unemployment.xls, Education.xls, PopulationEstimates.xls, DemoData.csv, weather.csv) by state and county into one data file in R. Sample codes are given below. Make sure your state and country have the same labels across all 7 data sets. Again, **the highlighted areas are placeholders** that will need to be customized, depending on how you name your variables and your data.

```
yourcombineddata<- merge(coviddata, poverty, by=c("state", "county"))
yourcombineddata<- merge(yourcombineddata, unemployment, by=c("state", "county"))
yourcombineddata<- merge(yourcombineddata, education, by=c("state", "county"))
yourcombineddata<- merge(yourcombineddata, population, by=c("state", "county"))
yourcombineddata<- merge(yourcombineddata, demo, by=c("state", "county"))
yourcombineddata<- merge(yourcombineddata, weather, by=c("state", "county"))
yourcombineddata$Med_HH_Income_2018 <- as.numeric(gsub('\\$|,', '',
yourcombineddata$Med_HH_Income_2018))
```

### Analysis

- Run the following R codes (Generalized Linear mixed/GLMM effect models with offset "county-level population", random effects on the intercept and slope with county as the independent unit) to identify predictors that statistically significantly associated with covid-19 incidence (**the highlighted areas are placeholders** that will need to be customized, depending on how you name the variables in your own data; the variables that are not highlighted have their default labels, unless you change yourself).

```
library("lme4")
stdz<-function(x) (x-mean(x))/sd(x) # standardize the predictors
case.model<- glmer(cases ~ state+stdz(Page1)+stdz(Page2)+stdz(Page3)+stdz(Page4)
+stdz(Page5)+stdz(Page6)+stdz(Page7)+stdz(Page8)+stdz(Page9)
+stdz(Page10)+stdz(Page11)+stdz(Page12)+stdz(Page13)
+stdz(Page14)+stdz(Page15)+stdz(Page16)+stdz(Page17)
+stdz(Pmale)+stdz(Pwhite)+stdz(Pblack)
+stdz(Pasian)+stdz(Phispanic)+stdz(MaxTemp)+stdz(MinTemp)
+stdz(precipitation)+stdz(PpovertyALL_2018)
+stdz(Ppoverty017_2018)+stdz(Unemployment_rate_2018)
+stdz(Med_HH_Income_2018)+stdz(Med_HH_Income_vs_Total_2018)
+stdz(Pnohighschool1418)+stdz(Phighschool1418)
+stdz(Psomecollege1418)+stdz(Pcollege1418)
+(1+stdz(date) | county), nAGQ = 0L,
offset = TOT_POP/1000000, data = yourcombineddata,
family = poisson)
summary(case.model)
```

- Run the following R codes (GLMM models with offset “cases”) to identify predictors that statistically significantly associated with covid-19 mortality (the highlighted areas are placeholders that will need to be customized, depending on how you name the variables in your own data; the variables that are not highlighted have their default labels, unless you change yourself).

```
death.model<- glmer(deaths ~ state+stdz(Page1)+stdz(Page2)+stdz(Page3)+stdz(Page4)
+stdz(Page5)+stdz(Page6)+stdz(Page7)+stdz(Page8)+stdz(Page9)
+stdz(Page10)+stdz(Page11)+stdz(Page12)+stdz(Page13)
+stdz(Page14)+stdz(Page15)+stdz(Page16)+stdz(Page17)
+stdz(Pmale)+stdz(Pwhite)+stdz(Pblack)
+stdz(Pasian)+stdz(Phispanic)+stdz(MaxTemp)+stdz(MinTemp)
+stdz(precipitation)+stdz(PpovertyALL_2018)
+stdz(Ppoverty017_2018)+stdz(Unemployment_rate_2018)
+stdz(Med_HH_Income_2018)+stdz(Med_HH_Income_vs_Total_2018)
+stdz(Pnohighschool1418)+stdz(Phighschool1418)
+stdz(Psomecollege1418)+stdz(Pcollege1418)
+(1+stdz(date) | county), nAGQ = 0L,
offset = (cases+0.5)/1000000, data = yourcombineddata,
family = poisson)
summary(death.model)
```

## Submission

1. The R codes for data processing and data analysis
2. The R outputs from the GLMM models
3. One paragraph that summarizes the statistical significance of the predictors for covid-19 incidence rate and mortality rate. Note the predictors in the models are standardized.
4. Either one of the following two
  - a. The data file (could be in .csv, .txt, xlsx formats) that contains the combined data from the 7 sources to which the models are fitted
  - b. The 'DemoData.csv' file+the 'weather.csv' file(s)