Final Project

Nick Milikich

R Codes and Outputs

Step 1:

```
covid = read.csv("covid19.txt", header = TRUE)
covid = covid[covid$state %in% c("Arizona", "New Mexico", "Texas"), ]
covid$state = as.character(covid$state)
covid$state[covid$state == "Arizona"] = "AZ"
covid$state[covid$state == "New Mexico"] = "NM"
covid$state[covid$state == "Texas"] = "TX"

covid$date = as.character(covid$date)
covid$date = as.Date(covid$date) - as.Date("2020/1/1")
```

Step 2-5:

```
poverty = data.frame(read_excel("PovertyEstimates.xls"))
poverty = poverty[poverty$Stabr %in% c("AZ", "NM", "TX"), ]
poverty = rename(poverty, state = Stabr, county = Area_name)
poverty$county[substr(poverty$county, nchar(poverty$county) - 5,
  nchar(poverty$county)) == "County"] =
  substr(poverty$county[substr(poverty$county, nchar(poverty$county)
  - 5, nchar(poverty$county)) == "County"], 1,
  nchar(poverty$county[substr(poverty$county, nchar(poverty$county)
  - 5, nchar(poverty$county)) == "County"]) - 7)

unemployment = data.frame(read_excel("Unemployment.xls"))
unemployment = unemployment[unemployment$State %in% c("AZ", "NM", "TX"), ]
unemployment = rename(unemployment, state = State, county = Area_name)
unemployment$county[substr(unemployment$county,
  nchar(unemployment$county) - 9, nchar(unemployment$county) - 4) ==
  "County"] = substr(unemployment$county[substr(unemployment$county,
  nchar(unemployment$county) - 9, nchar(unemployment$county) - 4) ==
  "County"], 1, nchar(unemployment$county[substr(unemployment$county,
  nchar(unemployment$county) - 9, nchar(unemployment$county) - 4) ==
  "County"]) - 11)

education = data.frame(read_excel("Education.xls"))
education = education[education$State %in% c("AZ", "NM", "TX"), ]
education = rename(education, state = State, county = Area.name)
education$county[substr(education$county, nchar(education$county) - 5,
  nchar(education$county)) == "County"] =
  substr(education$county[substr(education$county, nchar(education$county)
  - 5, nchar(education$county)) == "County"], 1,
  nchar(education$county[substr(education$county, nchar(education$county)
```

```
  - 5, nchar(education$county)) == "County"]) - 7)

population = data.frame(read_excel("PopulationEstimates.xls"))
population = population[population$State %in% c("AZ", "NM", "TX"), ]
population = rename(population, state = State, county = Area_Name)
population$county[substr(population$county, nchar(population$county) - 5,
  nchar(population$county)) == "County"] =
  substr(population$county[substr(population$county,
  nchar(population$county) - 5, nchar(population$county)) == "County"], 1,
  nchar(population$county[substr(population$county, nchar(population$county)
  - 5, nchar(population$county)) == "County"]) - 7)
```

Step 6:

```
demo0<-read.csv("ArizonaDemographics.csv",header=T)
demo1<-read.csv("NewMexicoDemographics.csv",header=T)
demo2<-read.csv("TexasDemographics.csv",header=T)
demo0<-rbind(demo0,demo1,demo2)
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))
demoData<-cbind(total[c(1, 2, 4)], Pmale, Pwhite, Pblack, Pasian,
Phispanic, Page);

demo = demoData
demo = rename(demo, state = STNAME, county = CTYNAME)
demo$county = as.character(demo$county)
demo$state = as.character(demo$state)
demo$county[23] = "Dona Ana County"
demo$county = substr(demo$county, 1, nchar(demo$county) - 7)
demo$state[demo$state == "Arizona"] = "AZ"
demo$state[demo$state == "New Mexico"] = "NM"
demo$state[demo$state == "Texas"] = "TX"
```

Step 7:

```
az.max = read.csv("ArizonaMaxTemp.csv", skip = 3, header = TRUE)
az.min = read.csv("ArizonaMinTemp.csv", skip = 3, header = TRUE)
az.prc = read.csv("ArizonaPrecip.csv", skip = 3, header = TRUE)
az = cbind("AZ", as.character(az.max$Location), az.max$Value, az.min$Value, az.prc$Value)
colnames(az) = c("state", "county", "Max Temp", "Min Temp", "Precip")

nm.max = read.csv("NewMexicoMaxTemp.csv", skip = 3, header = TRUE)
```

```
nm.min = read.csv("NewMexicoMinTemp.csv", skip = 3, header = TRUE)
nm.prc = read.csv("NewMexicoPrecip.csv", skip = 3, header = TRUE)
nm = cbind("NM", as.character(nm.max$Location), nm.max$Value, nm.min$Value, nm.prc$Value)
colnames(nm) = c("state", "county", "Max Temp", "Min Temp", "Precip")

tx.max = read.csv("TexasMaxTemp.csv", skip = 3, header = TRUE)
tx.min = read.csv("TexasMinTemp.csv", skip = 3, header = TRUE)
tx.prc = read.csv("TexasPrecip.csv", skip = 3, header = TRUE)
tx = cbind("TX", as.character(tx.max$Location), tx.max$Value, tx.min$Value, tx.prc$Value)
colnames(tx) = c("state", "county", "Max Temp", "Min Temp", "Precip")

weather = data.frame(rbind(az, nm, tx))
weather$state = as.character(weather$state)
weather$county = as.character(weather$county)

weather$county[substr(weather$county, nchar(weather$county) - 5,
  nchar(weather$county)) == "County"] =
  substr(weather$county[substr(weather$county, nchar(weather$county) - 5,
  nchar(weather$county)) == "County"], 1,
  nchar(weather$county[substr(weather$county, nchar(weather$county) - 5,
  nchar(weather$county)) == "County"]) - 7)
```

Merging data by state and county:

```
yourcombineddata<- merge(covid, 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))

yourcombineddata$state = as.factor(yourcombineddata$state)
yourcombineddata$Max.Temp = as.numeric(yourcombineddata$Max.Temp)
yourcombineddata$Min.Temp = as.numeric(yourcombineddata$Min.Temp)
yourcombineddata$Precip = as.numeric(yourcombineddata$Precip)

write.csv(x = yourcombineddata, file = "combineddata.csv", row.names = FALSE)
```

Analysis:

```
library(lme4)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
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(Max.Temp) + stdz(Min.Temp) + stdz(Precip) + 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)
```

```
## Generalized linear mixed model fit by maximum likelihood (Adaptive
##   Gauss-Hermite Quadrature, nAGQ = 0) [glmerMod]
##  Family: poisson  ( log )
## Formula:
## 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(Max.Temp) + stdz(Min.Temp) + stdz(Precip) +
##     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)
##    Data: yourcombineddata
##  Offset: TOT_POP/1e+06
##
##      AIC      BIC   logLik deviance df.resid
##  35902.5  36160.6 -17911.2  35822.5     4646
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -20.0850  -0.5577  -0.0564   0.3801  15.7462
##
## Random effects:
##  Groups Name        Variance Std.Dev. Corr
##  county (Intercept) 0.6282   0.7926
##         stdz(date)  0.9509   0.9751   0.53
## Number of obs: 4686, groups:  county, 228
##
## Fixed effects:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      1.766999   0.287652   6.143 8.11e-10 ***
## stateNM                         -0.083378   0.556545  -0.150  0.88091
## stateTX                         -0.443628   0.341555  -1.299  0.19400
## stdz(Page1)                     -0.164554   0.291598  -0.564  0.57254
## stdz(Page2)                      0.345631   0.304214   1.136  0.25590
## stdz(Page3)                      0.011003   0.277181   0.040  0.96833
## stdz(Page4)                      0.052215   0.265097   0.197  0.84385
## stdz(Page5)                     -0.241119   0.444995  -0.542  0.58793
## stdz(Page6)                      0.488451   0.337487   1.447  0.14781
## stdz(Page7)                     -0.372265   0.297755  -1.250  0.21121
```

```
## stdz(Page8)                          0.473283   0.229428   2.063  0.03912 *
## stdz(Page9)                         -0.134850   0.216382  -0.623  0.53315
## stdz(Page10)                         0.073769   0.204381   0.361  0.71814
## stdz(Page11)                        -0.187001   0.163921  -1.141  0.25395
## stdz(Page12)                         0.157545   0.259214   0.608  0.54333
## stdz(Page13)                        -0.067760   0.355243  -0.191  0.84873
## stdz(Page14)                         0.243415   0.464404   0.524  0.60018
## stdz(Page15)                         0.296398   0.435879   0.680  0.49650
## stdz(Page16)                        -0.007199   0.427686  -0.017  0.98657
## stdz(Page17)                        -0.355962   0.355414  -1.002  0.31657
## stdz(Pmale)                         -0.195388   0.096037  -2.034  0.04190 *
## stdz(Pwhite)                        -0.303056   0.116844  -2.594  0.00950 **
## stdz(Pblack)                        -0.042097   0.089704  -0.469  0.63886
## stdz(Pasian)                        -0.026394   0.101053  -0.261  0.79394
## stdz(Phispanic)                      0.022630   0.126428   0.179  0.85794
## stdz(Max.Temp)                      -0.311023   0.269694  -1.153  0.24881
## stdz(Min.Temp)                       0.476839   0.272974   1.747  0.08067 .
## stdz(Precip)                        -0.198119   0.086158  -2.299  0.02148 *
## stdz(PpovertyALL_2018)               0.662316   0.284349   2.329  0.01985 *
## stdz(Ppoverty017_2018)              -1.016122   0.277438  -3.663  0.00025 ***
## stdz(Unemployment_rate_2018)         0.091827   0.078406   1.171  0.24153
## stdz(Med_HH_Income_2018)             0.576731   0.612364   0.942  0.34629
## stdz(Med_HH_Income_vs_Total_2018) -1.092967   0.578692  -1.889  0.05893 .
## stdz(Pnohighschool1418)            16.172730   6.356145   2.544  0.01095 *
## stdz(Phighschool1418)              14.323387   5.690996   2.517  0.01184 *
## stdz(Psomecollege1418)             10.493078   4.133814   2.538  0.01114 *
## stdz(Pcollege1418)                 21.977304   8.381707   2.622  0.00874 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Correlation matrix not shown by default, as p = 37 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it
```

```
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(Max.Temp) + stdz(Min.Temp) + stdz(Precip) + 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)
```

```
## Generalized linear mixed model fit by maximum likelihood (Adaptive
##   Gauss-Hermite Quadrature, nAGQ = 0) [glmerMod]
##  Family: poisson  ( log )
## Formula:
## 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(Max.Temp) + stdz(Min.Temp) + stdz(Precip) +
##     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)
##    Data: yourcombineddata
##  Offset: (cases + 0.5)/1e+06
##
##      AIC      BIC   logLik deviance df.resid
##   4790.2   5048.3  -2355.1   4710.2     4646
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.29256 -0.19838 -0.06100 -0.03278  3.12726
##
## Random effects:
##  Groups Name        Variance Std.Dev. Corr
##  county (Intercept) 11.13    3.336
##         stdz(date)   2.58    1.606    0.90
## Number of obs: 4686, groups:  county, 228
##
## Fixed effects:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -1.41725    1.04708  -1.354  0.17588
## stateNM                         -2.66501    2.65057  -1.005  0.31468
## stateTX                         -3.48653    1.29592  -2.690  0.00714 **
## stdz(Page1)                      0.52307    1.53747   0.340  0.73369
## stdz(Page2)                     -0.22528    1.56423  -0.144  0.88548
## stdz(Page3)                      0.11299    1.36467   0.083  0.93401
## stdz(Page4)                      0.21854    1.34698   0.162  0.87111
## stdz(Page5)                     -0.22637    2.24832  -0.101  0.91980
## stdz(Page6)                      1.53151    1.65596   0.925  0.35504
## stdz(Page7)                     -1.02771    1.48281  -0.693  0.48826
## stdz(Page8)                      1.18526    1.21371   0.977  0.32878
## stdz(Page9)                     -1.07076    1.22240  -0.876  0.38106
## stdz(Page10)                     0.92962    1.09946   0.846  0.39782
## stdz(Page11)                    -0.35802    0.94611  -0.378  0.70512
## stdz(Page12)                     0.41344    1.41194   0.293  0.76966
## stdz(Page13)                     1.51412    1.78678   0.847  0.39677
## stdz(Page14)                     0.33610    2.39646   0.140  0.88846
## stdz(Page15)                    -2.83553    2.40902  -1.177  0.23918
## stdz(Page16)                     2.07162    2.44748   0.846  0.39731
## stdz(Page17)                    -0.64020    1.85649  -0.345  0.73021
## stdz(Pmale)                     -0.72833    0.49006  -1.486  0.13723
## stdz(Pwhite)                     0.39455    0.50406   0.783  0.43377
## stdz(Pblack)                     0.47723    0.37644   1.268  0.20490
## stdz(Pasian)                     0.04242    0.35730   0.119  0.90549
## stdz(Phispanic)                 -0.34208    0.56663  -0.604  0.54604
## stdz(Max.Temp)                   0.41688    1.09323   0.381  0.70296
## stdz(Min.Temp)                   0.36872    1.07842   0.342  0.73242
## stdz(Precip)                     0.26746    0.32431   0.825  0.40955
```

```
## stdz(PpovertyALL_2018)             0.67638    1.15945   0.583  0.55965
## stdz(Ppoverty017_2018)             0.04543    1.19552   0.038  0.96969
## stdz(Unemployment_rate_2018)      -0.41199    0.35104  -1.174  0.24055
## stdz(Med_HH_Income_2018)          -1.04316    2.83817  -0.368  0.71321
## stdz(Med_HH_Income_vs_Total_2018)  1.47362    2.73531   0.539  0.59007
## stdz(Pnohighschool1418)            0.78827   24.79938   0.032  0.97464
## stdz(Phighschool1418)              0.21860   22.28086   0.010  0.99217
## stdz(Psomecollege1418)             0.27831   16.20758   0.017  0.98630
## stdz(Pcollege1418)                 1.32788   32.77848   0.041  0.96769
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Correlation matrix not shown by default, as p = 37 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it
```

Summary

The first generalized linear mixed effects model is fit to identify predictors associated with Covid-19 incidence. Incidence is a measure of the number of cases per population, so the response variable is the number of cases, which is offset by the total population of the county. `Page8`, a variable reporting the proportion of the population in that county in a particular age bracket, is significant, but only marginally so, and its significance seems likely to be a type I error since there are 17 variables associated with age bracket percentages, none of the others are close to significant, and the number of variables makes it likely for one to be reported significant by chance. There are several demographic variables that are signifant, including `Pmale` (percentage of male residents), `Pwhite` (percentage of white residents), `PpovertyALL_2018` (percentage of people living in poverty), `Ppoverty017_2018` (percentage of people age 0-17 living in poverty), `Med_HH_Income_vs_Total_2018` (the median household income as a percentage of the state median household income), `Pnohighschool1418` (percentage of residents who did not attend any high school), `Phighschool1418` (percentage of residents whose highest level of education was a high school diploma), `Psomecollege1418` (percentage of residents who had attended college but not received a degree), and `Pcollege1418` (percentage of residents whose highest level of education was a college degree). The magnitude of the coefficients is difficult to interpret since the data are standardized, but this suggests that there is some correlation between socioeconomic status and the incidence of Covid-19: counties that are less white, less male, more impoverished overall, and have lower incomes tend to have a higher incidence of Covid-19. This analysis siggests that state and age distribution are not relevant to Covid-19 incidence. Finally, two variables related to weather, `Min.Temp` (monthly minimum temperature) and `Precip` (monthly precipitation) are significant. Counties that are warmer (have a higher minimum temperature) and have less precipitation tend to experience a higher incidence rate of Covid-19, and it would be worth investigating further if this is simply due to a localized demographic distribution in Texas, New Mexico, and Arizona that happens to correlate with weather patterns, or if this is a more widespread trend.

The second generalized linear mixed effects model is fit to identify predictors associated with Covid-19 death rate, a measure of the number of deaths per cases, so the response variable is the number of deaths, which is offset by the total number of cases in the county (with a continuity correction of 0.5 since some counties have as few as 1 case). Far fewer covariates are significant in predicting death rate: the only significant predictor is the state being Texas, which suggests that the death rate in Texas is significantly lower than the death rate in Arizona. This is probably due to Texas and Arizona being in different stages of outbreak for most of the available data, but could be due to policy or medical differences between the states; it is worth investigating how these death rates compare to the rest of the coutry. Overall, it seems that socioeconomic data, age and ethnicity distribution, education, and weather data have little relevance to Covid-19 death rate. It would be reasonable to expect that socioeconomic differences could lead to differential ability to treat Covid-19 and lead to differing death rates, but it seems that the disease affects people roughly equally once they have caught it.