Statistical Methods for Discrete Response, Time Series, and Panel Data (W271): Group Lab 3

Due 4pm Tuesday August 11 2020

```
# clear cache in R
rm(list = ls())
```

```
# libraries
library(foreign)
library(gplots)
library(ggplot2)
library(stats)
library(Hmisc)
library(car)
library(usmap)
library(dplyr)
library(gridExtra)
library(stargazer)
```

U.S. traffic fatalities: 1980-2004

In this lab, you are asked to answer the question **"Do changes in traffic laws affect traffic fatalities?"** To do so, you will conduct the tasks specified below using the data set *driving.Rdata*, which includes 25 years of data that cover changes in various state drunk driving, seat belt, and speed limit laws.

Specifically, this data set contains data for the 48 continental U.S. states from 1980 through 2004. Various driving laws are indicated in the data set, such as the alcohol level at which drivers are considered legally intoxicated. There are also indicators for "per se" laws—where licenses can be revoked without a trial—and seat belt laws. A few economics and demographic variables are also included. The description of the each of the variables in the dataset is come with the dataste.

Exercises:

1. (30%) Load the data. Provide a description of the basic structure of the dataset, as we have done throughout the semester. Conduct a very thorough EDA, which should include both graphical and tabular techniques, on the dataset, including both the dependent variable *totfatrte* and the potential explanatory variables. You need to write a detailed narrative of your observations of your EDA. *Reminder: giving an "output dump" (i.e. providing a bunch of graphs and tables without description and hoping your audience will interpret them) will receive a zero in this exercise.*

```
# load the dataset
load("driving.RData")
#str(data)
#desc

# one row per year per state
head(table(data$year, data$state))
```

```
##
##        1 3 4 5 6 7 8 10 11 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
##   1980 1 1 1 1 1 1 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##   1981 1 1 1 1 1 1 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##   1982 1 1 1 1 1 1 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##   1983 1 1 1 1 1 1 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##   1984 1 1 1 1 1 1 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##   1985 1 1 1 1 1 1 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##
##        30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51
##   1980  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##   1981  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##   1982  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##   1983  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##   1984  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##   1985  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
```

```
max(data$year)
```

```
## [1] 2004
```

There are 1,200 observations and 56 variables in this dataset. This is a panel dataset of *state* (51 states) and *year* (1980-2004). *totoatrte* is the dependent variable in this study. Based on the variable descriptions, there are a number of potential explanatory variables as follow:
* 6 speed limit variables – *sl55*, *sl65*, *sl70*, *sl75*, *slnone*, *sl70plus*
* Seatbelt – *seatbelt*
* Minimum drinking age – *minage*
* Zero tolerance law – *zerotol*
* Graduated drivers license law – *gdl*
* Admin license recovation (per se law) – *perse*
* 2 Blood alcohol limit variables – *bac10*, *bac08*
* State population – *statepop*
* Unemployment rate – *unem*
* Percentage of population aged 14 through 24 – *perc14_24*
* 2 Seat belt laws – *sbprim*, sb_secon_
* 25 dummy variables represneting year 1980 to 2004 – *d80* ... *d04*
* Vehicle miles traveled per capita – *vehicmilespc*

```
# average fatality rate per 100,000 across states
state_avg <- data %>% group_by(state) %>% summarise(avg_totfatrte=mean(totfatrte))
```

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

```
fips_map <-read.csv("statecodes.csv")

state_avg2 <- merge(x=fips_map, y=state_avg, by="state",all.x = TRUE)[,c("code", "avg_totfatrte

state_avg2 <- rename(state_avg2, c("value"="avg_totfatrte", "state"="code"))

p1.1 <- plot_usmap(data = state_avg2,  values="value", color = "red") +
  scale_fill_continuous(name="", low="white", high="red") +
  theme(legend.position = "right") +ggtitle("Average fatality rate per 100,000 (1980-2004)")

# average state's fatality rate per 100,000 across years
year_avg <- data %>% group_by(year) %>% summarise(avg_totfatrte=mean(totfatrte))
```

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

```
# nation's fatality rate per 100,000 across years
total.rate.pop <- data %>% group_by(year) %>% summarise(total_rate = sum(totfat)*100000/sum(sta
```

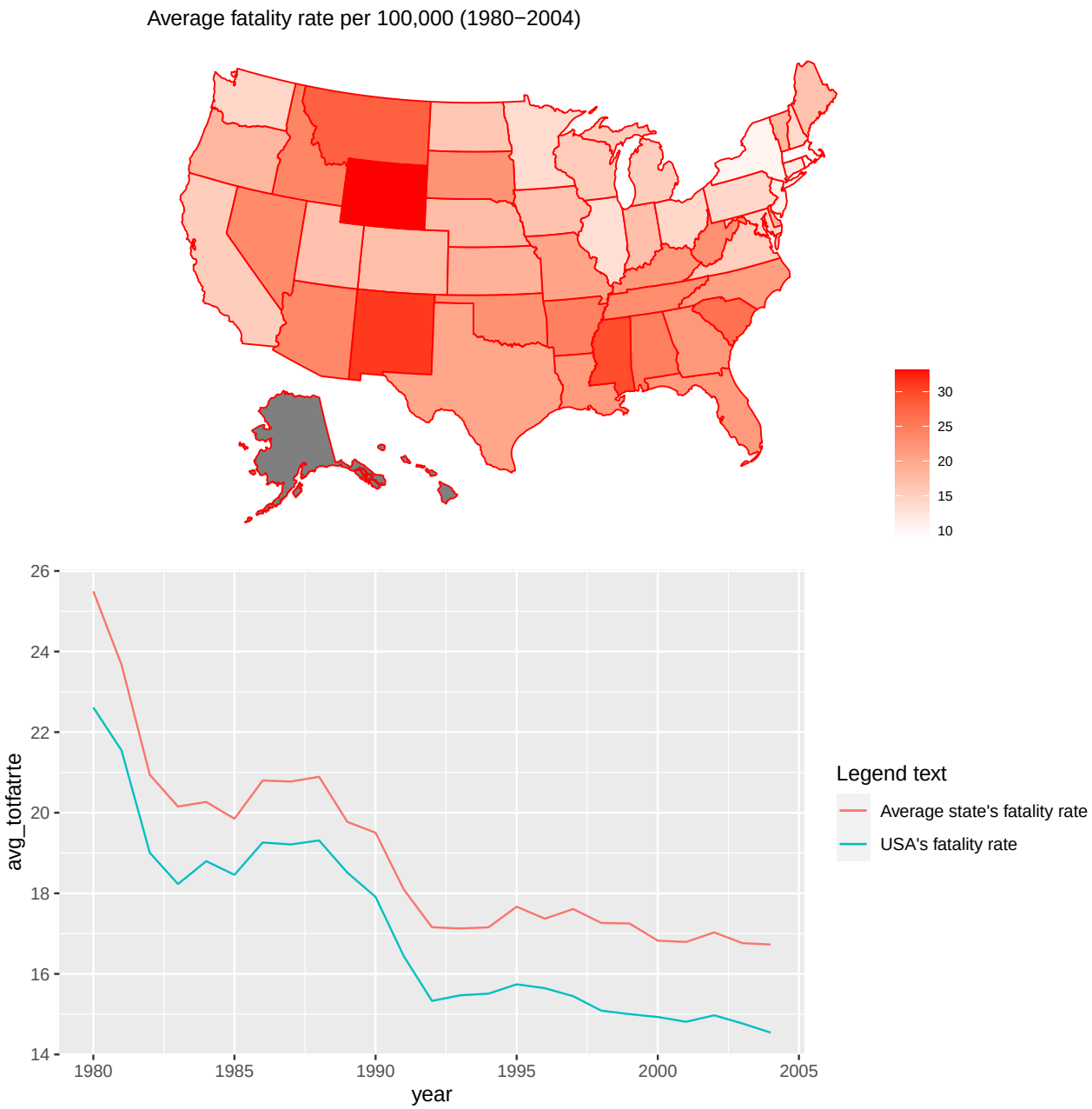
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
year_avg$totlfarte <- total.rate.pop$total_rate

# fatality trend over the years
p1.2 <- ggplot(year_avg, aes(x=year)) +
  geom_line(aes(y=avg_totfatrte, color="Average state's fatality rate")) +
  geom_line(aes(y=totlfarte, color="USA's fatality rate")) +
```

```
  labs(color="Legend text")

grid.arrange(p1.1, p1.2, nrow=2)
```



The plot above shows (i) average fatality rate (averaged over 25 years) of each state, and (ii) the fatality rates over the years (one plot is averaged across the states and the other is the entire national rate).

Western (Wyoming, New Mexico, etc.) and Southern states (Mississippi, etc.) show relatively high fatality rates compared to the rest of the country such as Midwest (Illinois, Minnesota, etc.) and coastal states (New York, California, etc.)

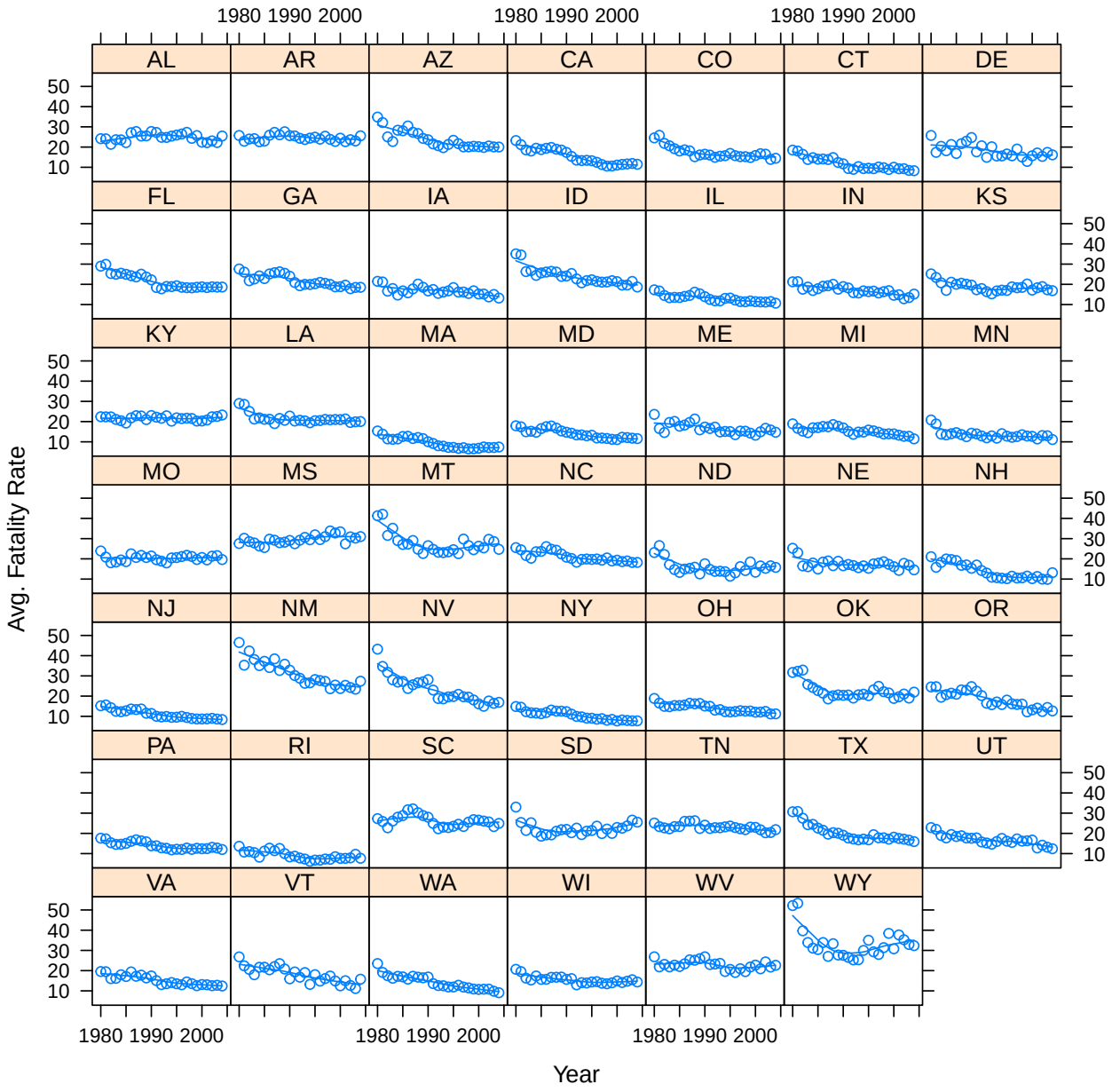
Over the 25 years, the fatality rate declined steeply in the early 80s and the early 90s, and has overall

a declining trend. Another interesting observation is that the national fatality rate consistently trails off the average across states. The higher average fatality rate (when averaging across states) suggests that some smaller states (i.e., fewer population) over weigh the average rate with relatively higher fatality rate.

Growth Gurve Analysis

- Note general flat to downward trend with exception of Mississippi
- Nevada and New Mexico drop looks steep

```
data_state <- merge(x=data, y=fips_map, by="state",all.x = TRUE)
xyplot(totfatrte~year | code, data=data_state,
       prepanel = function(x, y) prepanel.loess(x, y, family="gaussian"),
       xlab = "Year", ylab = "Avg. Fatality Rate",
       panel = function(x, y) {
          panel.xyplot(x, y)
          panel.loess(x,y, family="gaussian") },
       as.table=T)
```




```
# this is hard to read!
#g <- ggplot(data_state, aes(year, totfatrte, colour = as.factor(code)))
#g + geom_line() + ggtitle("Growth Curve by state")
```

Plan for the remaining EDA

Growth curve analysis - conditional color plot given various key variables: - seatbelt law - speed limit - driver license law - zero tolerance law - per se law - blood alcohol limit

General correlations (entire data (on fixed effects) as well as selected years) - probably using a scatterplotmatrix - fatality vs: - state population - minimum drinking age - unemployment rate - percentage of population aged 14 through 24 - vehicles miles traveled per capita (*vehicmilespc*)

2. (15%) How is the our dependent variable of interest *totfatrte* defined? What is the average of this variable in each of the years in the time period covered in this dataset? Estimate a linear regression model of *totfatrte* on a set of dummy variables for the years 1981 through 2004. What does this model explain? Describe what you find in this model. Did driving become safer over this period? Please provide a detailed explanation.

3. (15%) Expand your model in *Exercise 2* by adding variables *bac08, bac10, perse, sbprim, sbsecon, sl70plus, gdl, perc14_24, unem, vehicmilespc*, and perhaps *transformations of some or all of these variables*. Please explain carefully your rationale, which should be based on your EDA, behind any transformation you made. If no transformation is made, explain why transformation is not needed. How are the variables *bac8* and *bac10* defined? Interpret the coefficients on *bac8* and *bac10*. Do *per se laws* have a negative effect on the fatality rate? What about having a primary seat belt law? (Note that if a law was enacted sometime within a year the fraction of the year is recorded in place of the zero-one indicator.)

4. (15%) Reestimate the model from *Exercise 3* using a fixed effects (at the state level) model. How do the coefficients on *bac08, bac10, perse, and sbprim* compare with the pooled OLS estimates? Which set of estimates do you think is more reliable? What assumptions are needed in each of these models? Are these assumptions reasonable in the current context?

5. (10%) Would you perfer to use a random effects model instead of the fixed effects model you built in *Exercise 4*? Please explain.

6. (10%) Suppose that *vehicmilespc*, the number of miles driven per capita, increases by $1,000$. Using the FE estimates, what is the estimated effect on *totfatrte*? Please interpret the estimate.

7. (5%) If there is serial correlation or heteroskedasticity in the idiosyncratic errors of the model, what would be the consequences on the estimators and their standard errors?