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

Due 4pm Tuesday August 11 2020

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

```r
library(foreign)
library(gplots)
library(ggplot2)
library(stats)
library(Hmisc)
library(car)
library(usmap)
library(dplyr)
library(gridExtra)
library(stargazer)
library(reshape2)
library(data.table)
library(tidyr)
library(grid)
library(plm)
```

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
```

```r
head(data, 10)
```

```
##    year state  sl55  sl65 sl70 sl75 slnone seatbelt minage zerotol gdl bac10
## 1  1980     1 1.000 0.000    0    0      0        0     18       0   0     1
## 2  1981     1 1.000 0.000    0    0      0        0     18       0   0     1
## 3  1982     1 1.000 0.000    0    0      0        0     18       0   0     1
## 4  1983     1 1.000 0.000    0    0      0        0     18       0   0     1
## 5  1984     1 1.000 0.000    0    0      0        0     18       0   0     1
## 6  1985     1 1.000 0.000    0    0      0        0     20       0   0     1
## 7  1986     1 1.000 0.000    0    0      0        0     21       0   0     1
## 8  1987     1 0.542 0.458    0    0      0        0     21       0   0     1
## 9  1988     1 0.000 1.000    0    0      0        0     21       0   0     1
## 10 1989     1 0.000 1.000    0    0      0        0     21       0   0     1
##    bac08 perse totfat nghtfat wkndfat totfatpvm nghtfatpvm wkndfatpvm statepop
## 1      0     0    940     422     236     3.200      1.437      0.803  3893888
## 2      0     0    933     434     248     3.350      1.558      0.890  3918520
## 3      0     0    839     376     224     2.810      1.259      0.750  3925218
## 4      0     0    930     397     223     3.000      1.281      0.719  3934109
## 5      0     0    932     421     237     2.830      1.278      0.720  3951834
## 6      0     0    882     358     224     2.510      1.019      0.637  3972527
## 7      0     0   1080     500     279     3.177      1.471      0.821  3991569
## 8      0     0   1111     499     300     2.970      1.334      0.802  4015261
## 9      0     0   1024     423     226     2.580      1.066      0.569  4023858
## 10     0     0   1029     418     247     2.520      1.024      0.605  4030229
##    totfatrte nghtfatrte wkndfatrte vehicmiles unem perc14_24 sl70plus sbprim
## 1      24.14      10.84       6.06   29.37500  8.8      18.9        0      0
## 2      24.07      11.08       6.33   27.85200 10.7      18.7        0      0
## 3      21.37       9.58       5.71   29.85765 14.4      18.4        0      0
## 4      23.64      10.09       5.67   31.00000 13.7      18.0        0      0
## 5      23.58      10.65       6.00   32.93286 11.1      17.6        0      0
## 6      22.20       9.01       5.64   35.13944  8.9      17.3        0      0
## 7      27.08      12.53       6.99   33.99371  9.8      17.0        0      0
## 8      27.67      12.43       7.47   37.40741  7.8      16.6        0      0
## 9      25.45      10.51       5.62   39.68992  7.2      16.2        0      0
## 10     25.53      10.37       6.13   40.83333  7.0      15.8        0      0
##    sbsecon d80 d81 d82 d83 d84 d85 d86 d87 d88 d89 d90 d91 d92 d93 d94 d95 d96
## 1        0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## 2        0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## 3        0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## 4        0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0
## 5        0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0
## 6        0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0
## 7        0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0
## 8        0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0
## 9        0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0
## 10       0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0
```

```
##    d97 d98 d99 d00 d01 d02 d03 d04 vehicmilespc
## 1    0   0   0   0   0   0   0   0     7543.874
## 2    0   0   0   0   0   0   0   0     7107.785
## 3    0   0   0   0   0   0   0   0     7606.622
## 4    0   0   0   0   0   0   0   0     7879.802
## 5    0   0   0   0   0   0   0   0     8333.562
## 6    0   0   0   0   0   0   0   0     8845.614
## 7    0   0   0   0   0   0   0   0     8516.377
## 8    0   0   0   0   0   0   0   0     9316.308
## 9    0   0   0   0   0   0   0   0     9863.649
## 10   0   0   0   0   0   0   0   0    10131.764
```

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*, *sbsecon*
* 25 dummy variables represneting year 1980 to 2004 – *d80* ... *d04*
* Vehicle miles traveled per capita – *vehicmilespc*

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

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

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

d_data <- merge(x=data, y=fips_map, by="state",all.x = TRUE) %>% dplyr::select(-state)
d_data <- rename(d_data, c("state"="code"))
d_data <- data.table(d_data)

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)")
```

```r
# 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)
```

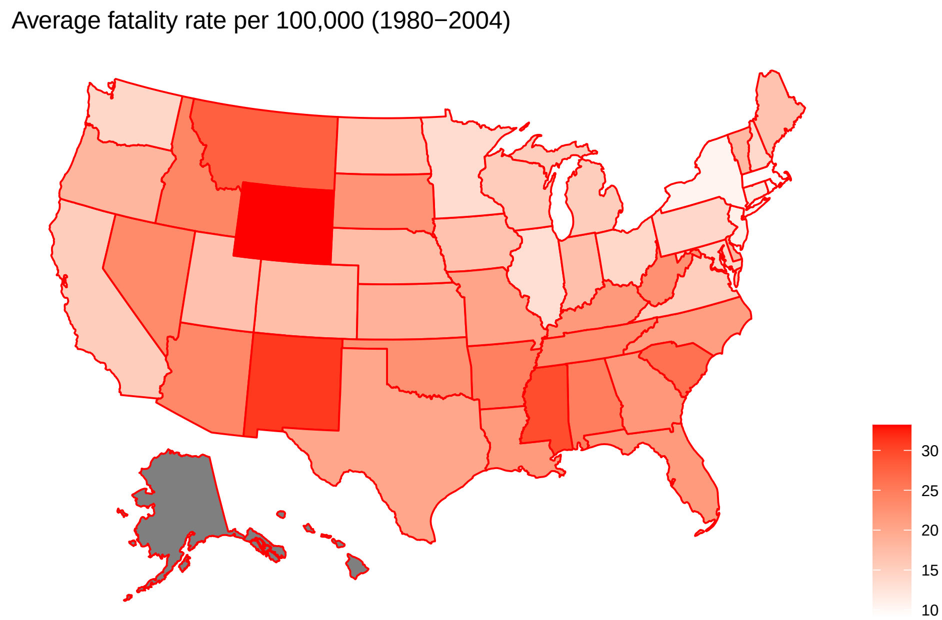
```r
# nation's fatality rate per 100,000 across years
total.rate.pop <- data %>% group_by(year) %>% summarise(total_rate = sum(totfat)*100000/sum(sta
```

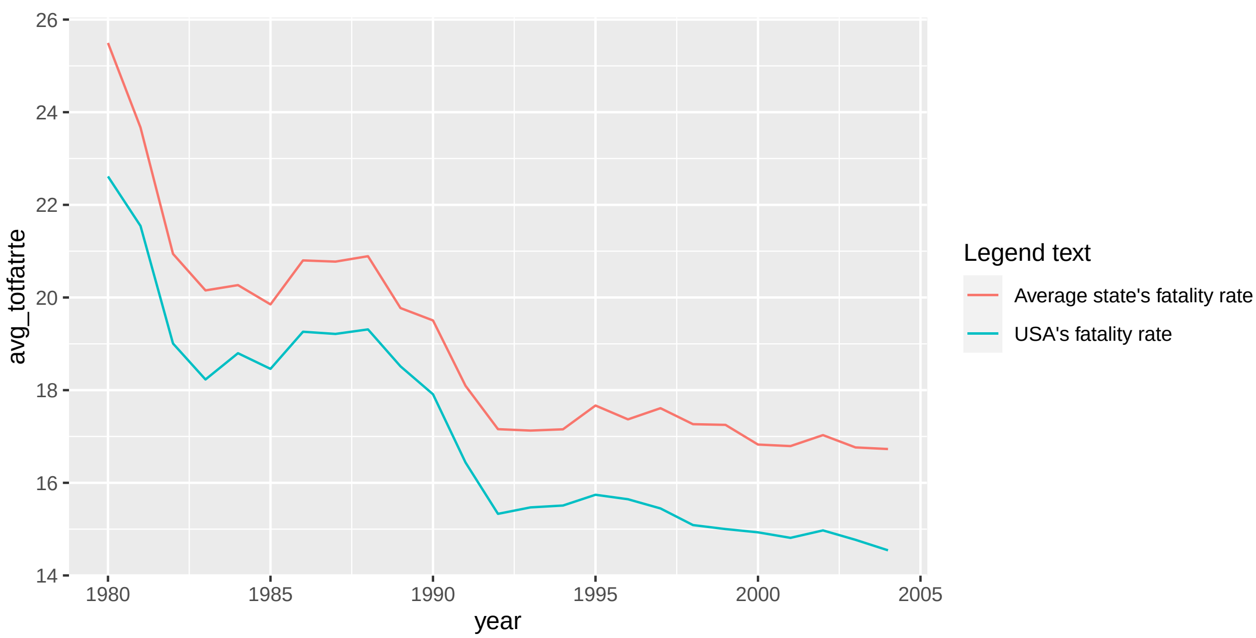
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```r
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 simple average of all the states and the other is the weighted average or the national rate).

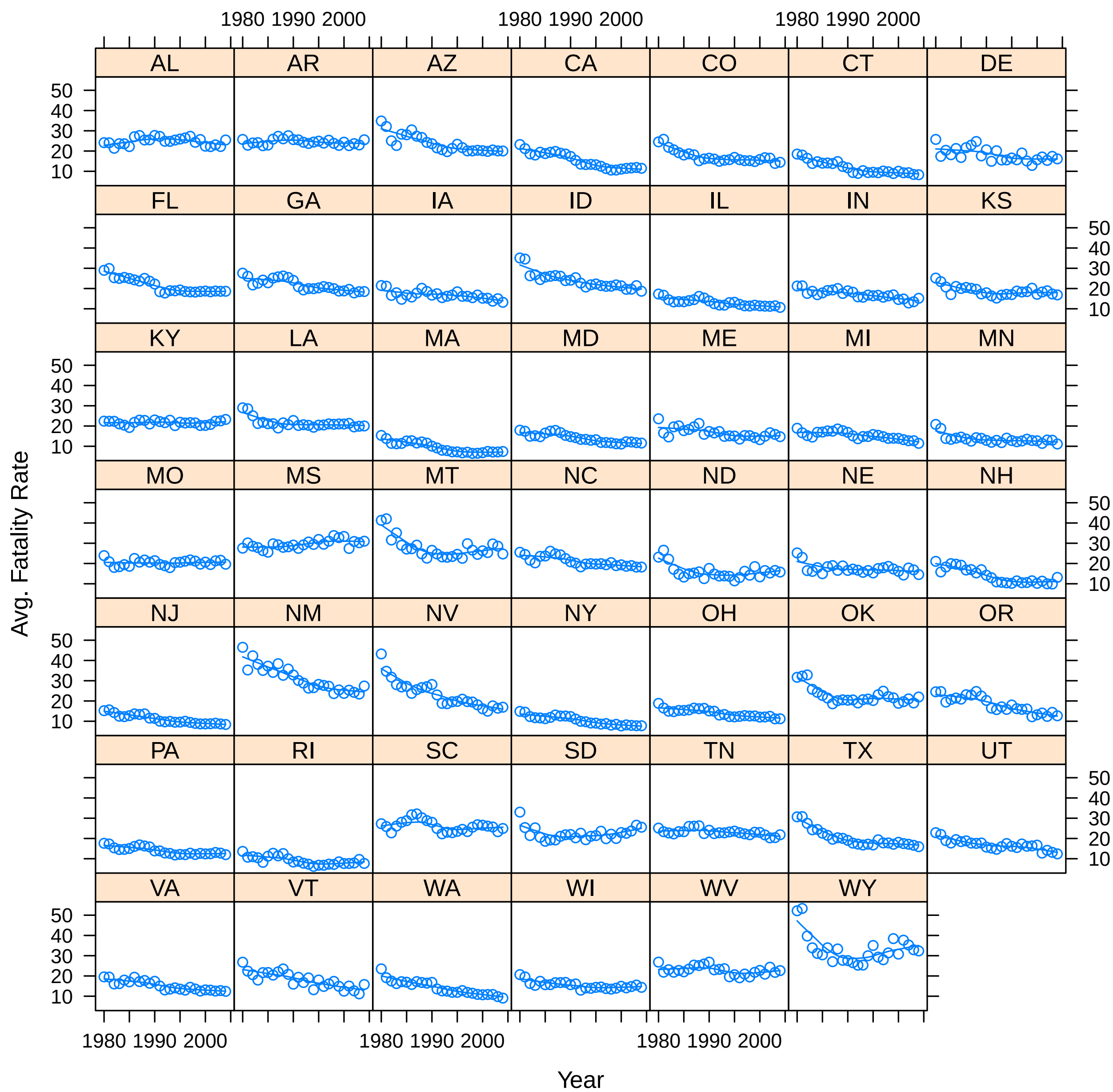
Western (Wyoming, New Mexico, etc.) and Southern states (Mississippi, etc.) show relatively high fatality rates compared to the rest of the country such as Midwest (Illinois, Minnesota, etc.) and coastal states (New York, California, etc.)

Over 25 years, the fatality rate declined steeply in the early 80s and the early 90s, and overall shows a persistent 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)
```



```r
# groups = carb < 2, col = c("blue", "red"))
```

Additional growth curves are plotted in Appendix I to show impact of certain shifts in laws and policies including speed limits, seat belt laws, minimum drinking age, zero tolerance law, graduated drivers license law, per se law, blood and alcohol limits. The key insights from these curves include:

(1) From Appendix I's figure 2 to figure 7, speed limits don't appear to be a key contributing factors on the fatality rate; the limit in most states were increased over the years, and yet the fatality rates kept gradually declining.

(2) From Appendix I's figure 8, the zero tolerance law seems to coincide with a sudden shift downwards or a persisting lower trend of the fatality rate in OH, OR, RI, NH, NJ, MA, AZ, and GA.

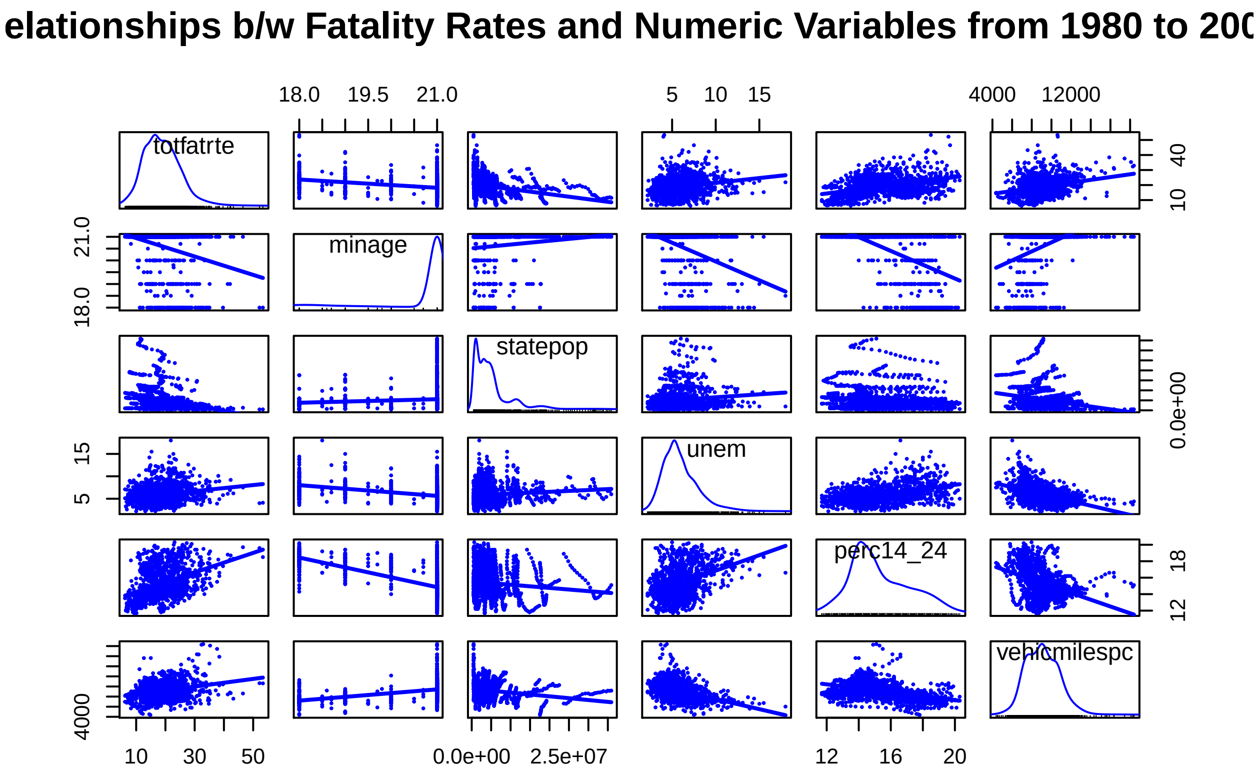
(3) From Appendix I's figure 9, no obvious pattern in relation between the fatality rate and the graduated license law; besides this law had recently been introduced in most states, so the data points may be not enough to draw its pattern/relationship with the fatality rate.

(4) From Appendix I's figure 10, the admin license revocation law seems to coincide with a sudden downward shift in the fatality rate in VT, OK, MN, ND, NH, LA, CA, CT, FL, and GA.

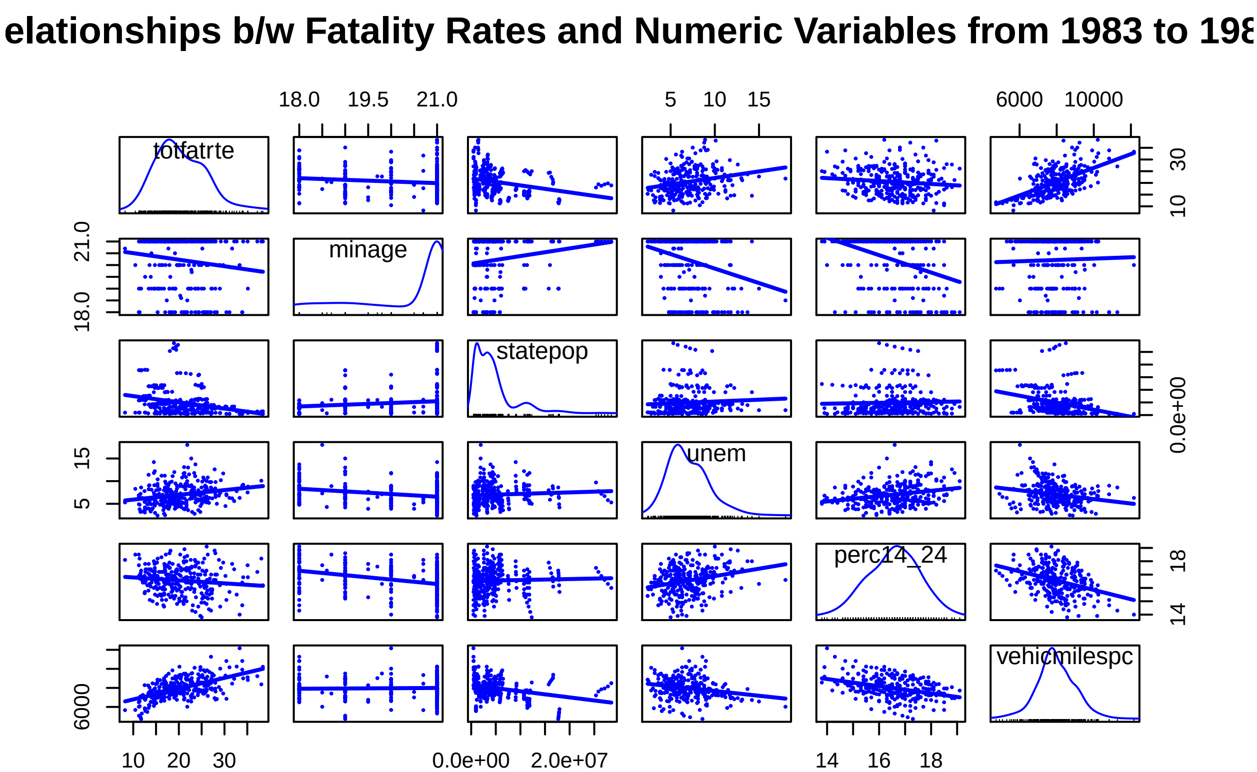
(5) From Appendix I'd figure 1 and figure 11 to 14, there is no obvious trend of alcohol limits and seat belt laws on changes in the fatality rate.

In addition to understanding these rules and policy impacts, next we plot the following scatterplot matrices to understand the relationships between fatality rate and other key state characteristics including minimum driving age, population, unemployment rate, percentage of teens, and miles traveled per capita.

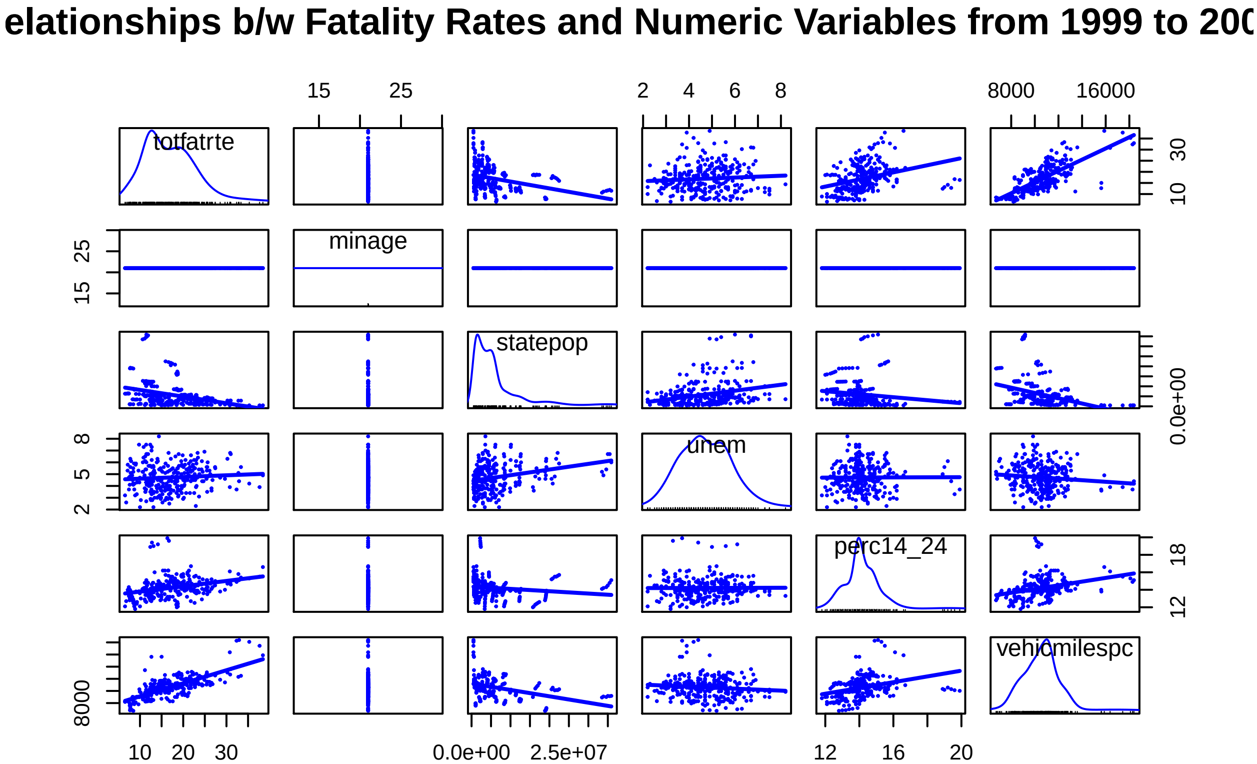
```r
# plot fatality rate and numeric variables scatterplots
scatterplotMatrix( ~ totfatrte + minage + statepop + unem + perc14_24 + vehicmilespc,
                   data = data_state, smooth= FALSE, cex= .2, diagonal=c("density"),
                   main='Relationships b/w Fatality Rates and Numeric Variables from 1980 to 20
```




```r
# plot fatality rate and numeric variables scatterplots
scatterplotMatrix( ~ totfatrte + minage + statepop + unem + perc14_24 + vehicmilespc,
                   data = data_state[data_state$year>=1983&data_state$year<=1988, ], smooth= F
                   main='Relationships b/w Fatality Rates and Numeric Variables from 1983 to 1
```



```r
# plot fatality rate and numeric variables scatterplots
scatterplotMatrix( ~ totfatrte + minage + statepop + unem + perc14_24 + vehicmilespc,
                    data = data_state[data_state$year>=1999&data_state$year<=2004, ], smooth= F
                    main='Relationships b/w Fatality Rates and Numeric Variables from 1999 to 20
```



Given the entire data set from 1980 to 2004, there are obvious positive relationships between (i) unemployment rate (ii) percentage of teens aged 14 to 24 and (iii) vehicle miles driven per capita. These relationships are directionally the same given the period of delcining fatality rate between 1999 and 2004. Interestingly, in the period when the fatality rate remained stubborn between 1983 and 1988, the relationship between the percentage of teen and the fatality rate was slightly negative; however, given a narrower percentage range (i.e., 14-19% compared with 12-20%, we may think the trend in this period was only temporal). Lastly, neither minimum driving age nor the state population seem to have a clear relationship with the fatality rate.

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.

Definition of the dependent variable and yearly average

The *totfatrte* is defined as the total fatality per 100,000 people. The averag in each year can be calculated in both simple average or weighted average, as shown in the table below.

```r
year_avg <- rename(year_avg,
                    c("simple average"="avg_totfatrte",
```

```r
                     "weighted average"="totlfarte"))

# yearly average nationwide
year_avg
```

```
## # A tibble: 25 x 3
##     year `simple average` `weighted average`
##    <int>            <dbl>              <dbl>
##  1  1980             25.5               22.6
##  2  1981             23.7               21.5
##  3  1982             20.9               19.0
##  4  1983             20.2               18.2
##  5  1984             20.3               18.8
##  6  1985             19.9               18.5
##  7  1986             20.8               19.3
##  8  1987             20.8               19.2
##  9  1988             20.9               19.3
## 10  1989             19.8               18.5
## # ... with 15 more rows
```

Regression model and explanation

This model gives us the time effect on the fatality rate. The intercept in this case is the simple average of *totfatrte* across all states in the omitted year 2004. Each of the coefficients `d80`, `d81`…`d04` is the average difference in *totfatrte* relative to the base year 2004.

```r
# regression model on the dummy year variables
stargazer(lm(totfatrte~d80+d81+d82+d83+d84+d85+d86+d87+d88+d89+d90+d91+d92+d93+d94+d95+d96+d97+
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                              totfatrte
## -----------------------------------------------
## d80                          8.766***
##                               (1.226)
##
## d81                          6.941***
##                               (1.226)
##
## d82                          4.214***
##                               (1.226)
##
## d83                          3.424***
##                               (1.226)
##
```

```
## d84                              3.539***
##                                  (1.226)
##
## d85                               3.123**
##                                  (1.226)
##
## d86                              4.071***
##                                  (1.226)
##
## d87                              4.046***
##                                  (1.226)
##
## d88                              4.163***
##                                  (1.226)
##
## d89                               3.043**
##                                  (1.226)
##
## d90                               2.776**
##                                  (1.226)
##
## d91                                1.366
##                                  (1.226)
##
## d92                                0.429
##                                  (1.226)
##
## d93                                0.399
##                                  (1.226)
##
## d94                                0.426
##                                  (1.226)
##
## d95                                0.940
##                                  (1.226)
##
## d96                                0.640
##                                  (1.226)
##
## d97                                0.882
##                                  (1.226)
##
## d98                                0.536
##                                  (1.226)
##
## d99                                0.521
##                                  (1.226)
##
```

```
## d00                               0.097
##                                  (1.226)
##
## d01                               0.064
##                                  (1.226)
##
## d02                               0.301
##                                  (1.226)
##
## d03                               0.035
##                                  (1.226)
##
## Constant                        16.729***
##                                  (0.867)
##
## -----------------------------------------------
## Observations                      1,200
## R2                                0.128
## Adjusted R2                       0.110
## Residual Std. Error      6.008 (df = 1175)
## F Statistic          7.164*** (df = 24; 1175)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

The outcome of the regression model suggests that (i) the base year 2004 has the average fatality rate of 16.73 per 100,000 people, (ii) the fatality rates before 2004 were all higher than that of 2004 with the average difference in each year as the coefficient of the respective dummy variable, and (iii) the average difference in the earlier years from 1980 to 1990 were statistically significant at p-value of 0.05; this is not surprising, given the standard error of the coefficient that is constant at 1.226 for all the dummy variables, implying that the larger the difference the more likely those will be statistically significant.

Given these outputs, we can say that the fatality rates for year 1991 to 2003 were not statistically different from the fatality rate in 2004; on the other hand, those for 1980 to 1990 were statistically different from that of 2004.

Did driving become safer

To answer this question, we should change the base year of this regression model to 1980, as opposed to 2004.

```r
stargazer(lm(totfatrte~factor(year), data=d_data), type = "text")
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                              totfatrte
## -----------------------------------------------
```

```
## factor(year)1981              -1.824
##                              (1.226)
##
## factor(year)1982             -4.552***
##                              (1.226)
##
## factor(year)1983             -5.342***
##                              (1.226)
##
## factor(year)1984             -5.227***
##                              (1.226)
##
## factor(year)1985             -5.643***
##                              (1.226)
##
## factor(year)1986             -4.694***
##                              (1.226)
##
## factor(year)1987             -4.720***
##                              (1.226)
##
## factor(year)1988             -4.603***
##                              (1.226)
##
## factor(year)1989             -5.722***
##                              (1.226)
##
## factor(year)1990             -5.989***
##                              (1.226)
##
## factor(year)1991             -7.400***
##                              (1.226)
##
## factor(year)1992             -8.337***
##                              (1.226)
##
## factor(year)1993             -8.367***
##                              (1.226)
##
## factor(year)1994             -8.339***
##                              (1.226)
##
## factor(year)1995             -7.826***
##                              (1.226)
##
## factor(year)1996             -8.125***
##                              (1.226)
##
```

```
## factor(year)1997            -7.884***
##                              (1.226)
##
## factor(year)1998            -8.229***
##                              (1.226)
##
## factor(year)1999            -8.244***
##                              (1.226)
##
## factor(year)2000            -8.669***
##                              (1.226)
##
## factor(year)2001            -8.702***
##                              (1.226)
##
## factor(year)2002            -8.465***
##                              (1.226)
##
## factor(year)2003            -8.731***
##                              (1.226)
##
## factor(year)2004            -8.766***
##                              (1.226)
##
## Constant                    25.495***
##                              (0.867)
##
## ----------------------------------------------
## Observations                  1,200
## R2                            0.128
## Adjusted R2                   0.110
## Residual Std. Error      6.008 (df = 1175)
## F Statistic          7.164*** (df = 24; 1175)
## ==============================================
## Note:              *p<0.1; **p<0.05; ***p<0.01
```

This model clearly highlights that driving has become safer between the years 1981 through 2004 with respect to the base year 1980 since the differences are significant at the < 0.05.

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.)

Variables *bac08, bac10, perse, sbprim, sbsecon, sl70plus, gdl and perc14_24* are yes-no indicator dummies, with the caveat that they can be fractional as noted in the problem statement. The fractional values will ideally need to be changed to 0 or 1 based on whether the variable is < 0.5 or >= 0.5. Note that we need to do special handling of the edge case of 0.5 in two categories. Variables `unem`, `perc14_24` and `vehiclesmilespc` are continuous. From the EDA we see that the distributions are normal and we do not see any reason for transformations. It may make sense to log transform `vehiclesmilespc` only if we want to interpret the coefficients in terms of percentage changes, but we leave it as is since we see no such requirement.

```
model.3 <- lm(totfatrte~bac08+bac10+perse+sbprim+sbsecon+sl70plus+gdl+perc14_24+unem+vehicmiles

stargazer(model.3, type="text")
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                              totfatrte
## ------------------------------------------------
## bac08                        -2.498***
##                               (0.538)
##
## bac10                        -1.418***
##                               (0.396)
##
## perse                        -0.620**
##                               (0.298)
##
## sbprim                        -0.075
##                               (0.491)
##
## sbsecon                        0.067
##                               (0.429)
##
## sl70plus                     3.348***
##                               (0.445)
##
## gdl                           -0.427
##                               (0.527)
##
## perc14_24                      0.142
##                               (0.123)
##
## unem                         0.757***
##                               (0.078)
##
## vehicmilespc                 0.003***
##                              (0.0001)
##
```

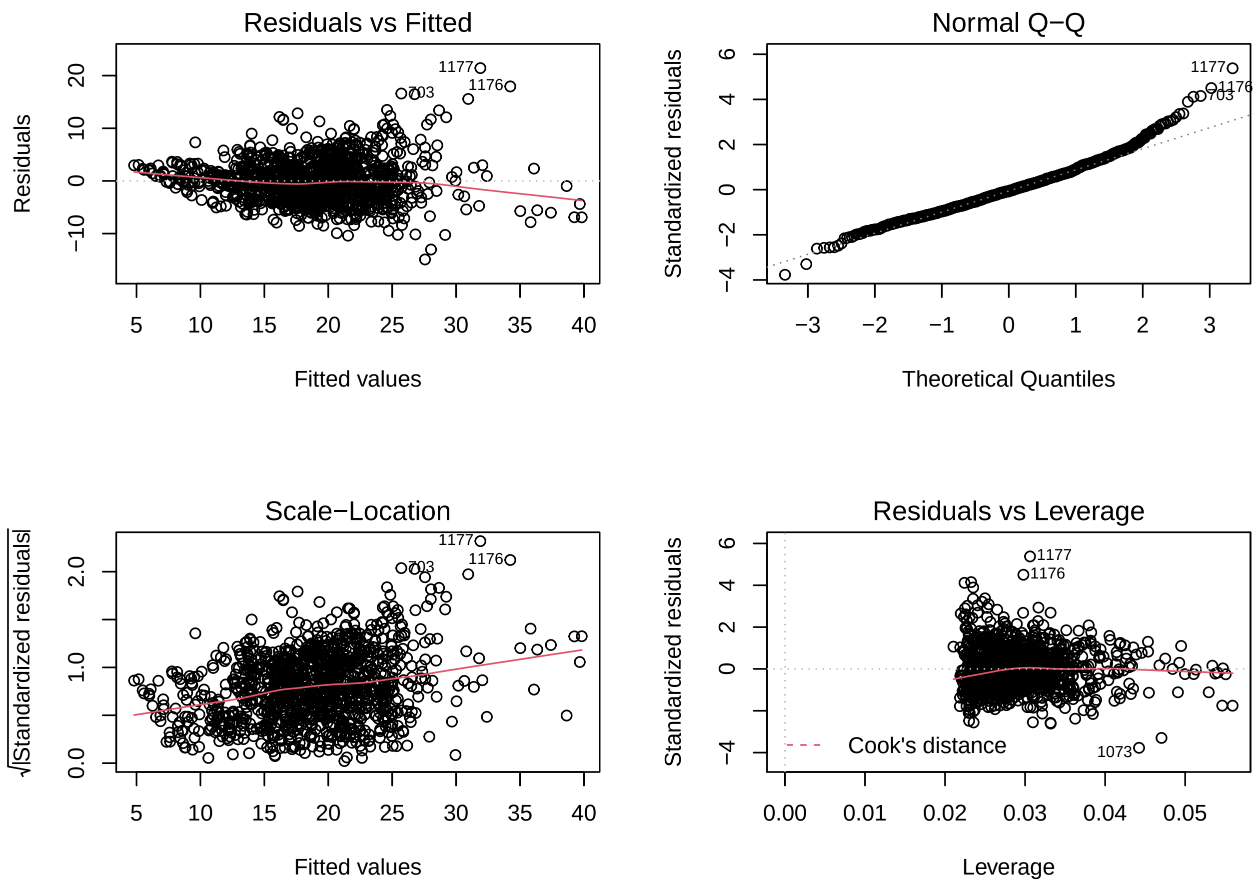
```
## factor(year)1981              -2.175***
##                               (0.828)
##
## factor(year)1982              -6.596***
##                               (0.853)
##
## factor(year)1983              -7.397***
##                               (0.869)
##
## factor(year)1984              -5.850***
##                               (0.876)
##
## factor(year)1985              -6.483***
##                               (0.895)
##
## factor(year)1986              -5.853***
##                               (0.931)
##
## factor(year)1987              -6.367***
##                               (0.967)
##
## factor(year)1988              -6.592***
##                               (1.014)
##
## factor(year)1989              -8.071***
##                               (1.053)
##
## factor(year)1990              -8.959***
##                               (1.077)
##
## factor(year)1991             -11.069***
##                               (1.101)
##
## factor(year)1992             -12.878***
##                               (1.123)
##
## factor(year)1993             -12.731***
##                               (1.136)
##
## factor(year)1994             -12.365***
##                               (1.157)
##
## factor(year)1995             -11.953***
##                               (1.184)
##
## factor(year)1996             -13.876***
##                               (1.223)
##
```

```
## factor(year)1997            -14.258***
##                              (1.250)
##
## factor(year)1998            -15.042***
##                              (1.265)
##
## factor(year)1999            -15.091***
##                              (1.284)
##
## factor(year)2000            -15.444***
##                              (1.305)
##
## factor(year)2001            -16.184***
##                              (1.334)
##
## factor(year)2002            -16.724***
##                              (1.348)
##
## factor(year)2003            -17.021***
##                              (1.359)
##
## factor(year)2004            -16.711***
##                              (1.387)
##
## Constant                     -2.716
##                              (2.476)
##
## -----------------------------------------------
## Observations                  1,200
## R2                            0.608
## Adjusted R2                   0.596
## Residual Std. Error      4.046 (df = 1165)
## F Statistic          53.096*** (df = 34; 1165)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

Check model assumptions

We see some evidence of heteroskedasticity pointing to omitted variables.

```r
par(mfrow=c(2,2))
plot(model.3)
```



Do *per se laws* have a negative effect on the fatality rate? From the model output above, it is evident that per se laws have had a negative effect on the fatality rate, as seen from the coefficient of -0.756 and p value < 0.05

What about having a primary seat belt law?

We do not see evidence of the primary seat belt law having any effect on the fatality rate as seen from the p value. This seems suspicous and points to the limitations inherent in pooled OLS models.

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?

The Pooled OLS model from Part 3 does not control for the fixed effect differences in each state. Each state has state specific factors that influence the fatality rate over time and the Pooled OLS model does not differentiate between the state specific differences over time and therefore treats all the observations the same way. Therefore the fixed effects model is more reliable for this inference. The output of the model produces different estimates from the pooled OLS model. We note that the model correctly shows the statistically significant negative effect of `sbprim` on the dependent variable.

```r
model.4 <-
  plm(
    totfatrte ~ bac08+bac10+perse+sbprim+sbsecon+sl70plus+gdl+perc14_24+unem+vehicmilespc+facto
    index = c("state"),
    model = "within",
    data = d_data
  )

stargazer(model.4, type="text")
```

```
##
## ============================================
##                      Dependent variable:
##                  ---------------------------
##                           totfatrte
## --------------------------------------------
## bac08                     -1.437***
##                            (0.394)
##
## bac10                     -1.063***
##                            (0.269)
##
## perse                     -1.152***
##                            (0.234)
##
## sbprim                    -1.227***
##                            (0.343)
##
## sbsecon                    -0.350
##                            (0.252)
##
## sl70plus                   -0.063
##                            (0.269)
##
## gdl                        -0.412
##                            (0.293)
##
## perc14_24                  0.187**
##                            (0.095)
##
## unem                      -0.572***
##                            (0.061)
##
## vehicmilespc              0.001***
##                           (0.0001)
##
## factor(year)1981          -1.511***
```

```
##                          (0.413)
##
## factor(year)1982        -3.025***
##                          (0.442)
##
## factor(year)1983        -3.504***
##                          (0.457)
##
## factor(year)1984        -4.259***
##                          (0.465)
##
## factor(year)1985        -4.727***
##                          (0.485)
##
## factor(year)1986        -3.661***
##                          (0.518)
##
## factor(year)1987        -4.306***
##                          (0.555)
##
## factor(year)1988        -4.767***
##                          (0.602)
##
## factor(year)1989        -6.130***
##                          (0.640)
##
## factor(year)1990        -6.230***
##                          (0.665)
##
## factor(year)1991        -6.917***
##                          (0.682)
##
## factor(year)1992        -7.774***
##                          (0.703)
##
## factor(year)1993        -8.094***
##                          (0.716)
##
## factor(year)1994        -8.504***
##                          (0.734)
##
## factor(year)1995        -8.255***
##                          (0.756)
##
## factor(year)1996        -8.607***
##                          (0.796)
##
## factor(year)1997        -8.708***
```

```
##                                (0.820)
##
## factor(year)1998             -9.349***
##                                (0.834)
##
## factor(year)1999             -9.475***
##                                (0.844)
##
## factor(year)2000             -9.992***
##                                (0.856)
##
## factor(year)2001             -9.631***
##                                (0.873)
##
## factor(year)2002             -8.907***
##                                (0.882)
##
## factor(year)2003             -8.937***
##                                (0.890)
##
## factor(year)2004             -9.339***
##                                (0.911)
##
## --------------------------------------------
## Observations                    1,200
## R2                              0.626
## Adjusted R2                     0.599
## F Statistic        55.094*** (df = 34; 1118)
## ============================================
## Note:              *p<0.1; **p<0.05; ***p<0.01
```

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

The null hypothesis of the Hausman test is the Random effects is preferred, the alternate hypothesis is that the fixed effects model is preferred. The below result suggests that the fixed effect model is preferred. A

```
model.re <-   plm(
    totfatrte ~ bac08+bac10+perse+sbprim+sbsecon+sl70plus+gdl+perc14_24+unem+vehicmilespc+facto
    index = c("state"),
    model = "random",
    data = d_data
  )
phtest(model.4, model.re)
```

```
##
##  Hausman Test
##
```

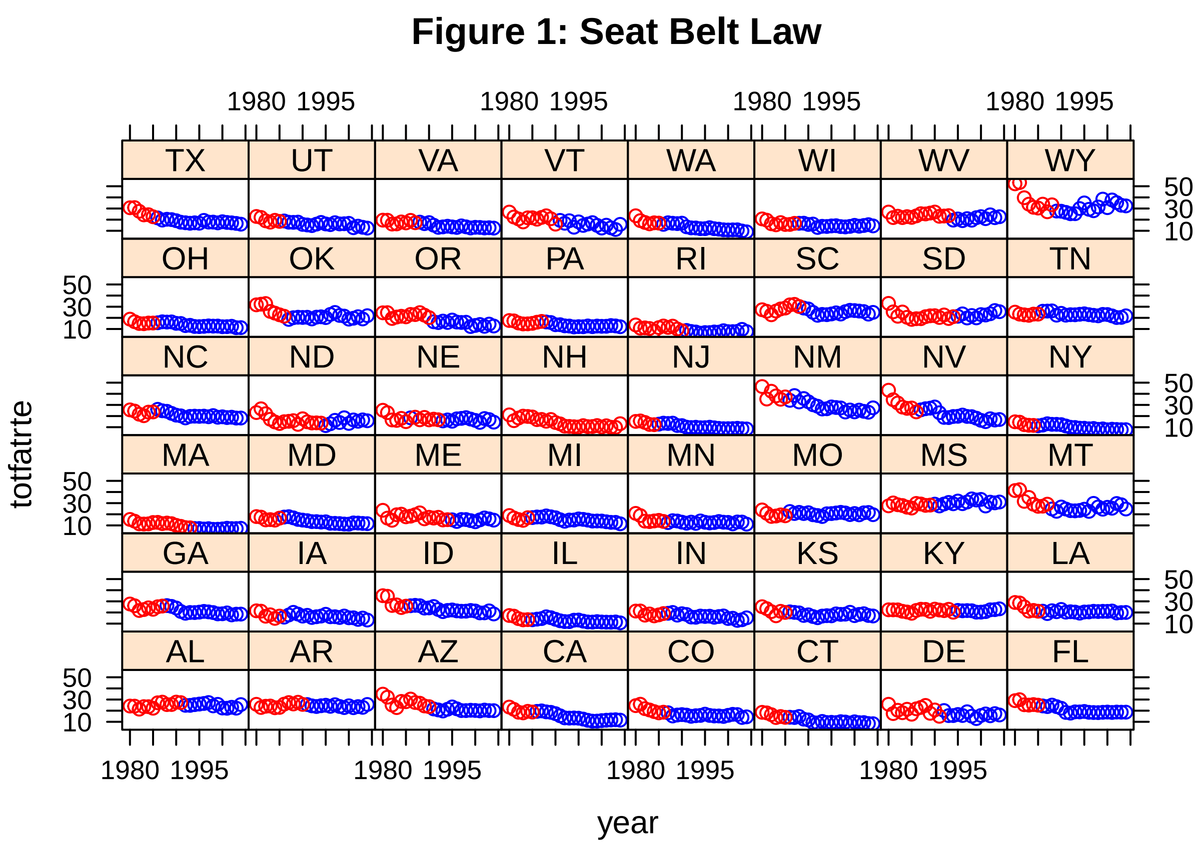
```
## data:  totfatrte ~ bac08 + bac10 + perse + sbprim + sbsecon + sl70plus +  ...
## chisq = 148.69, df = 34, p-value = 2.727e-16
## alternative hypothesis: one model is inconsistent
```

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?

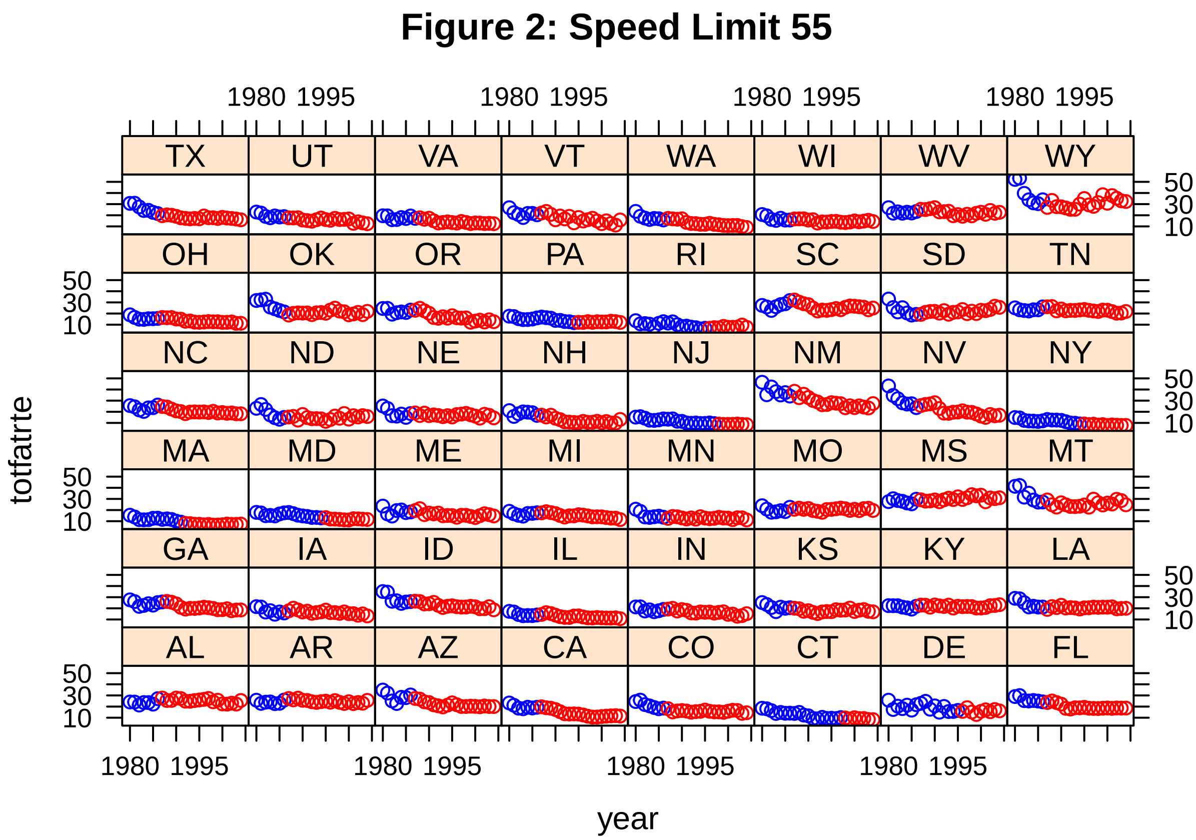
Appendix:

I. Growth Curve Analysis by certain key variables

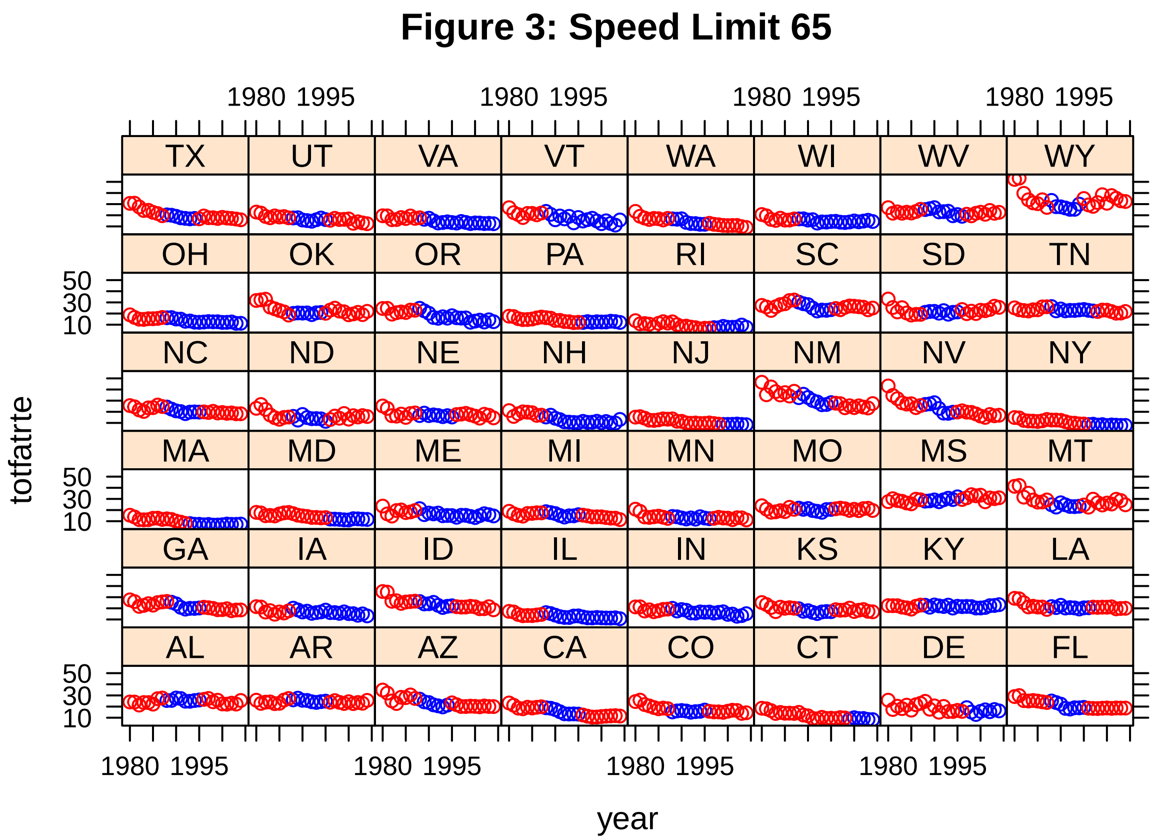
```
# seat belt law
xyplot(totfatrte~year | code, data=data_state,
       groups = seatbelt < 1, col = c("blue","red"),
       main = "Figure 1: Seat Belt Law")
```




```
# speed limit 55
xyplot(totfatrte~year | code, data=data_state,
       groups = sl55 < 1, col = c("blue","red"),
       main = "Figure 2: Speed Limit 55")
```

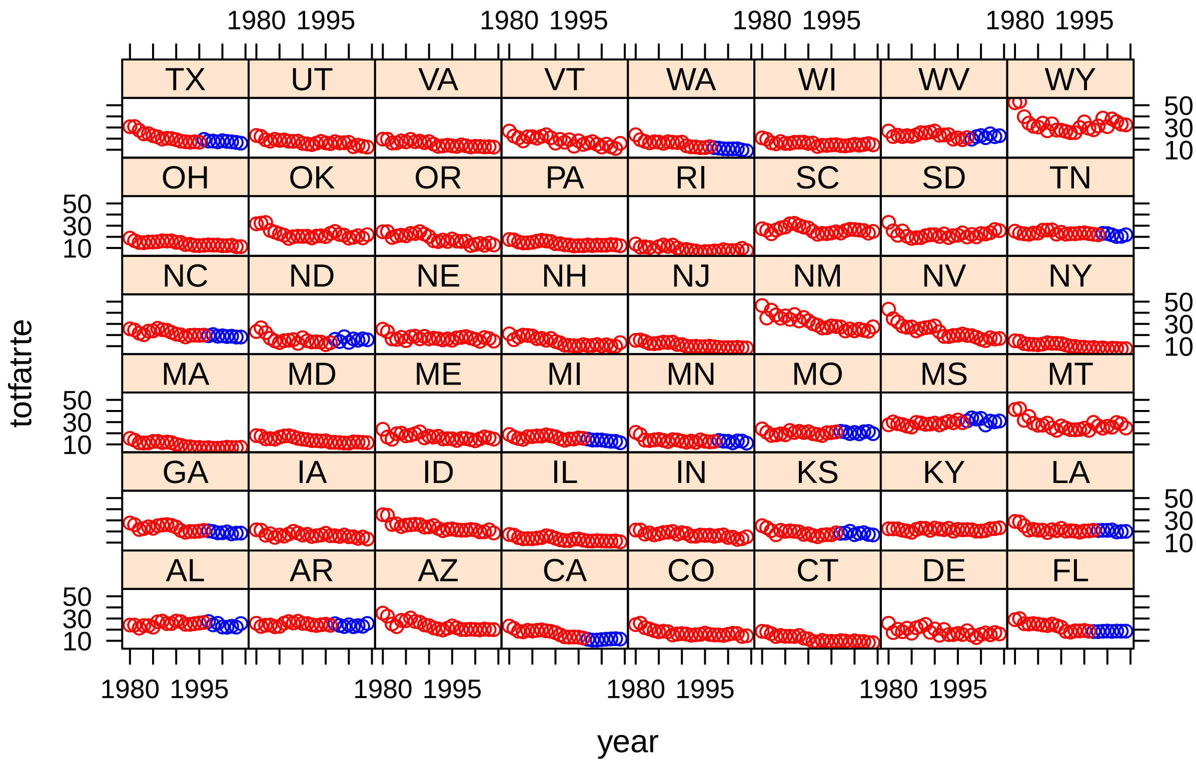



```r
# speed limit 65
xyplot(totfatrte~year | code, data=data_state,
       groups = sl65 < 1, col = c("blue","red"),
       main = "Figure 3: Speed Limit 65")
```

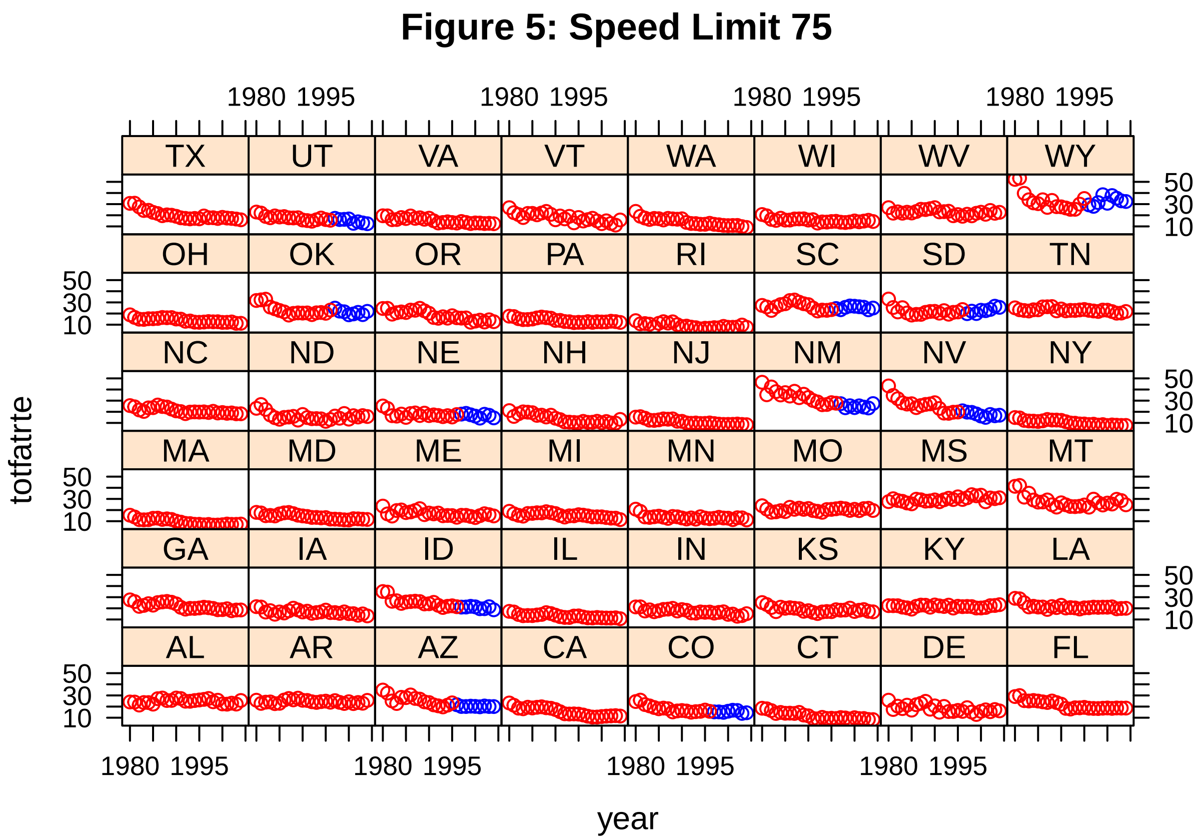



```r
# speed limit 70
xyplot(totfatrte~year | code, data=data_state,
       groups = sl70 < 1, col = c("blue","red"),
       main = "Figure 4: Speed Limit 70")
```

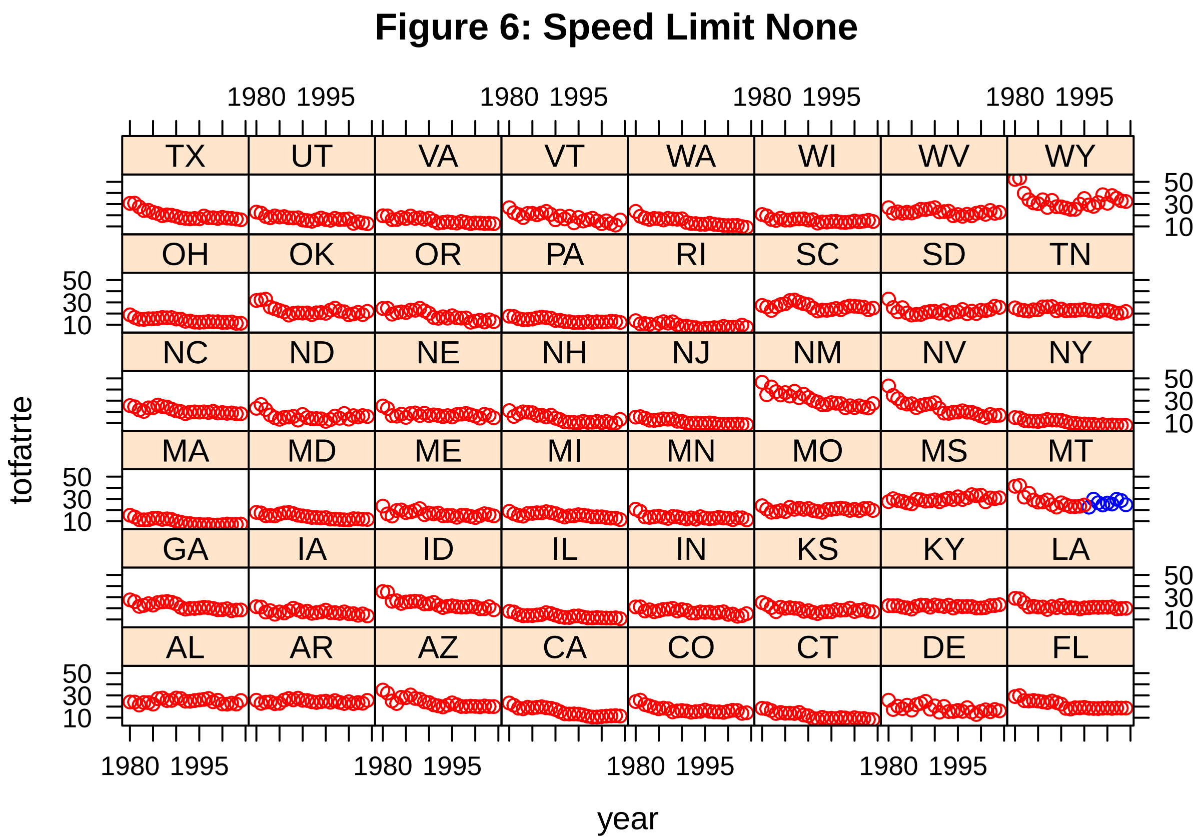
Figure 4: Speed Limit 70



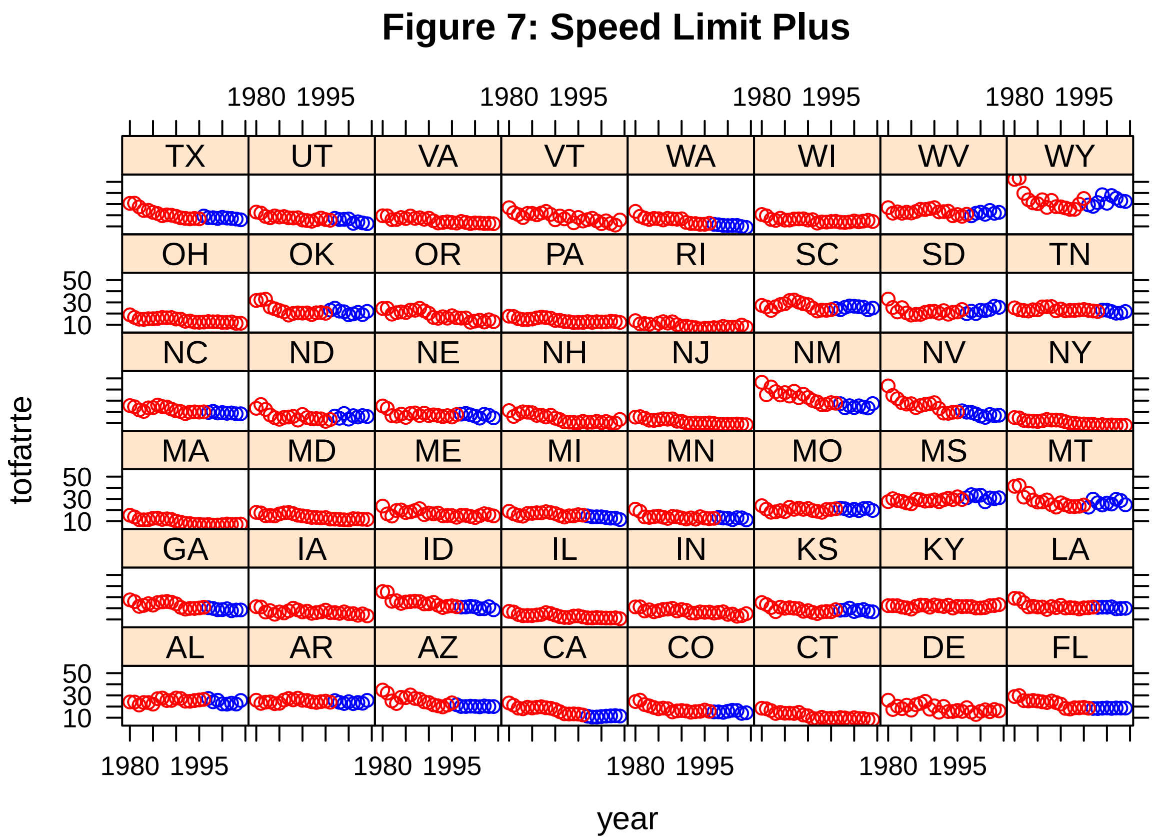

```r
# speed limit 75
xyplot(totfatrte~year | code, data=data_state,
       groups = sl75 < 1, col = c("blue","red"),
       main = "Figure 5: Speed Limit 75")
```




```
# speed limit slnone
xyplot(totfatrte~year | code, data=data_state,
       groups = slnone < 1, col = c("blue","red"),
       main = "Figure 6: Speed Limit None")
```

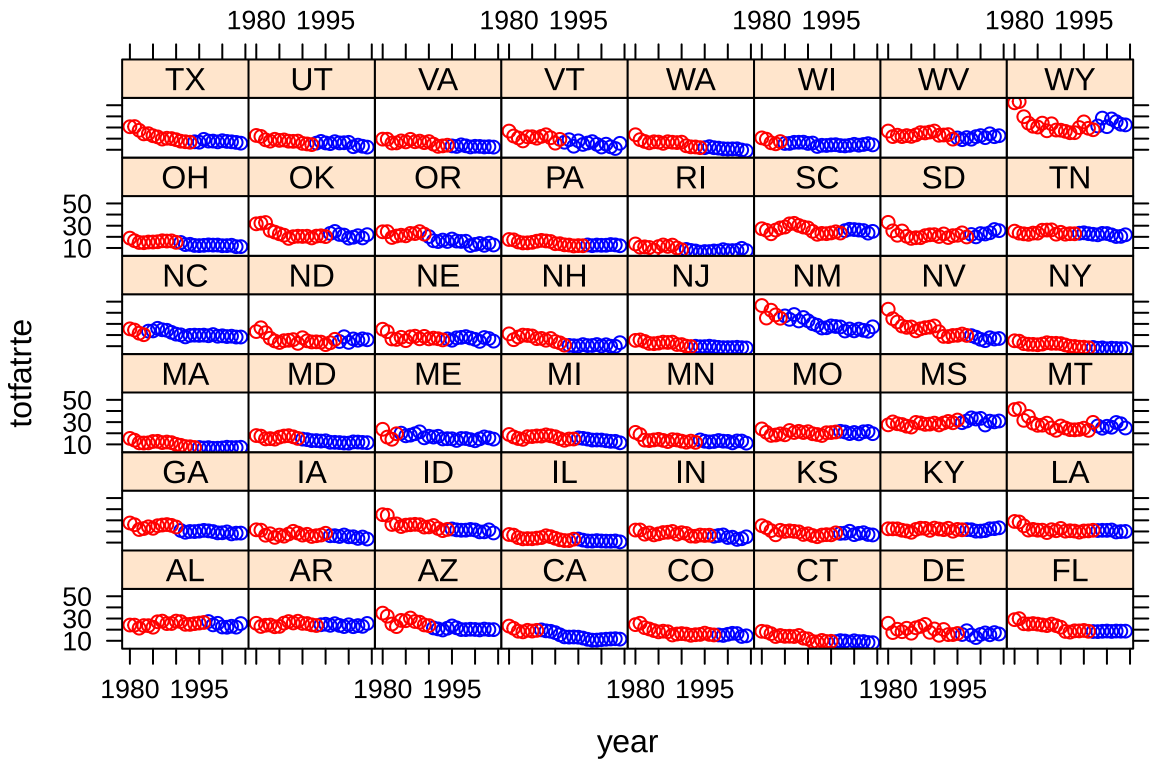



```r
# speed limit plus
xyplot(totfatrte~year | code, data=data_state,
       groups = sl70plus < 1, col = c("blue","red"),
       main = "Figure 7: Speed Limit Plus")
```

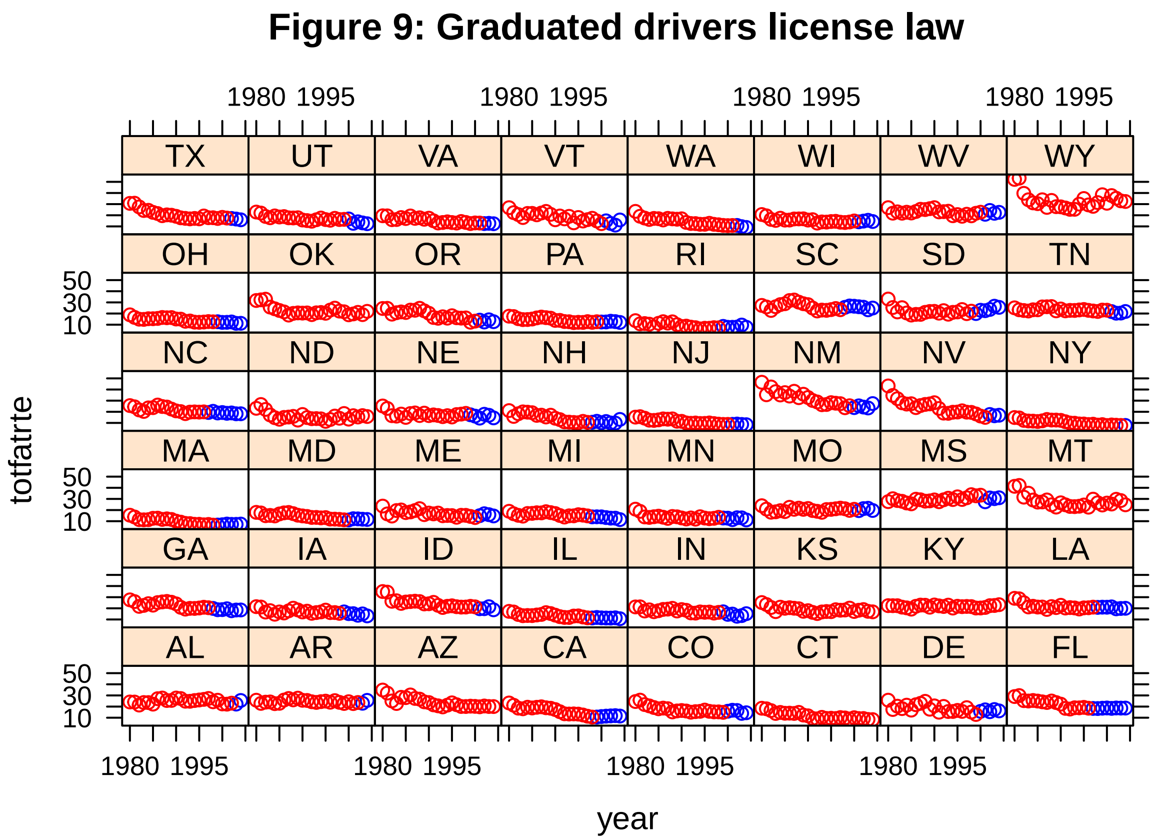



```
# zero tolerance law
xyplot(totfatrte~year | code, data=data_state,
       groups = zerotol < 1, col = c("blue","red"),
       main = "Figure 8: Zero tolerance law")
```

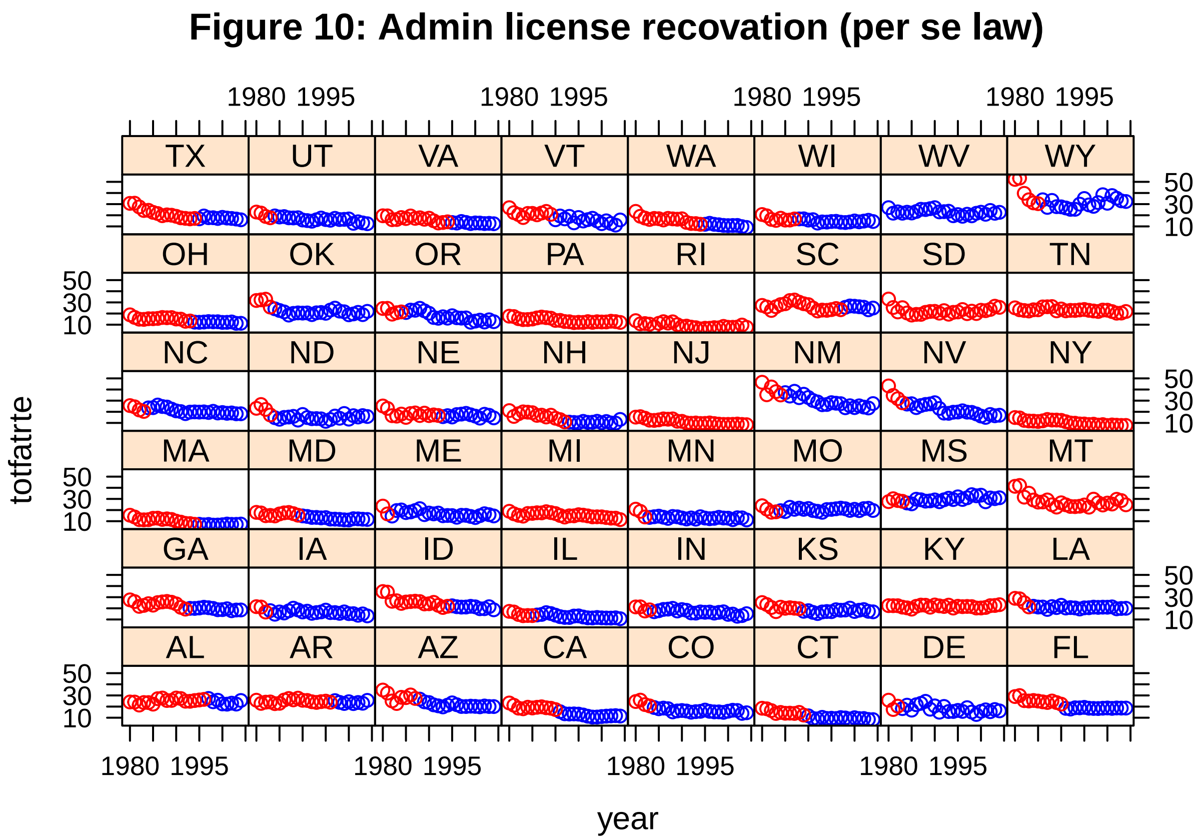
Figure 8: Zero tolerance law



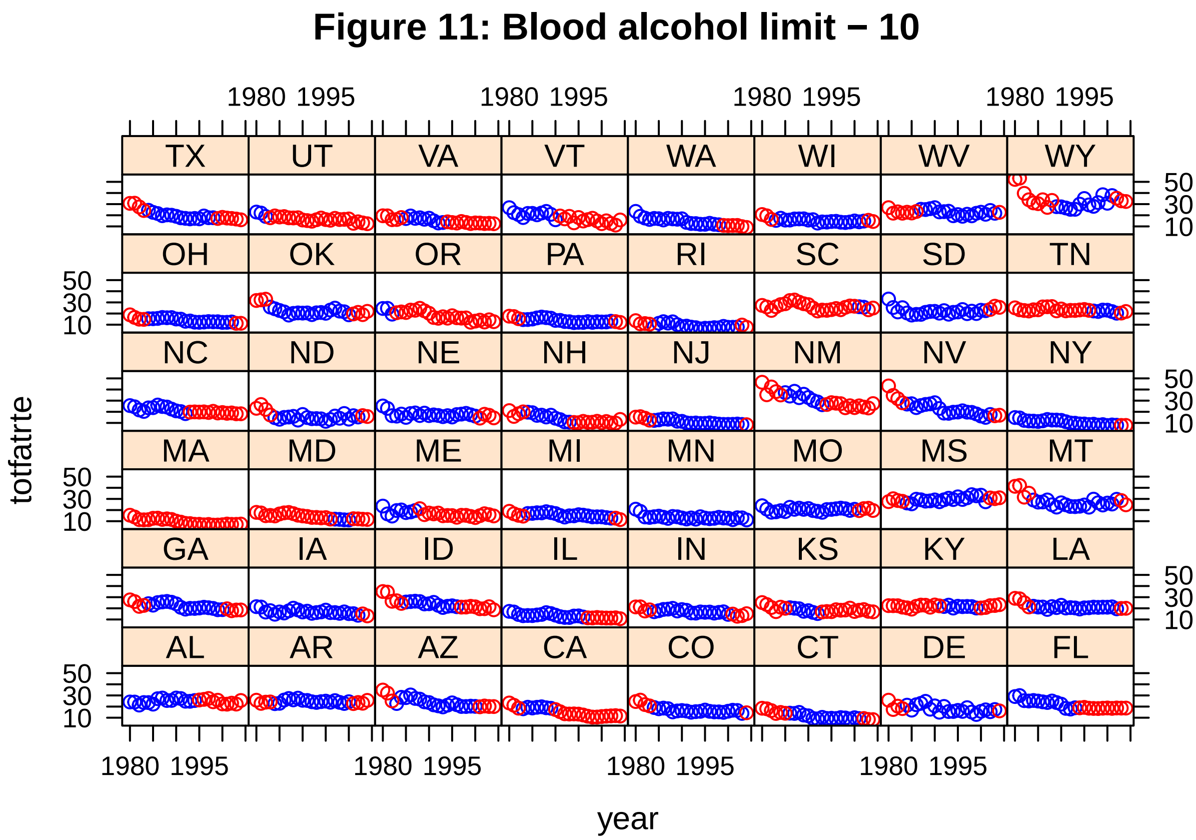

```r
# Graduated drivers license law
xyplot(totfatrte~year | code, data=data_state,
       groups = gdl < 1, col = c("blue","red"),
       main = "Figure 9: Graduated drivers license law")
```



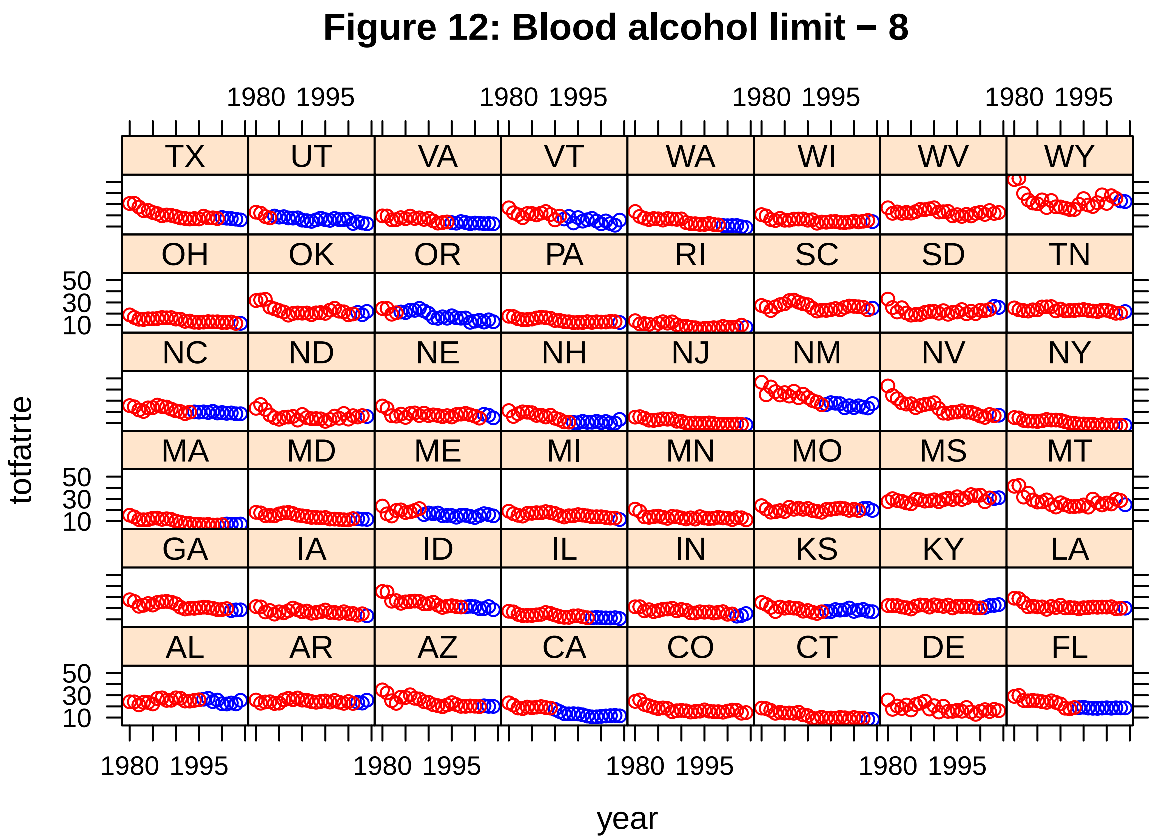

```
# Admin license recovation (per se law)
xyplot(totfatrte~year | code, data=data_state,
       groups = perse < 1, col = c("blue","red"),
       main = "Figure 10: Admin license recovation (per se law)")
```




```r
# Blood alcohol limit - 10
xyplot(totfatrte~year | code, data=data_state,
       groups = bac10 < 1, col = c("blue","red"),
       main = "Figure 11: Blood alcohol limit - 10")
```

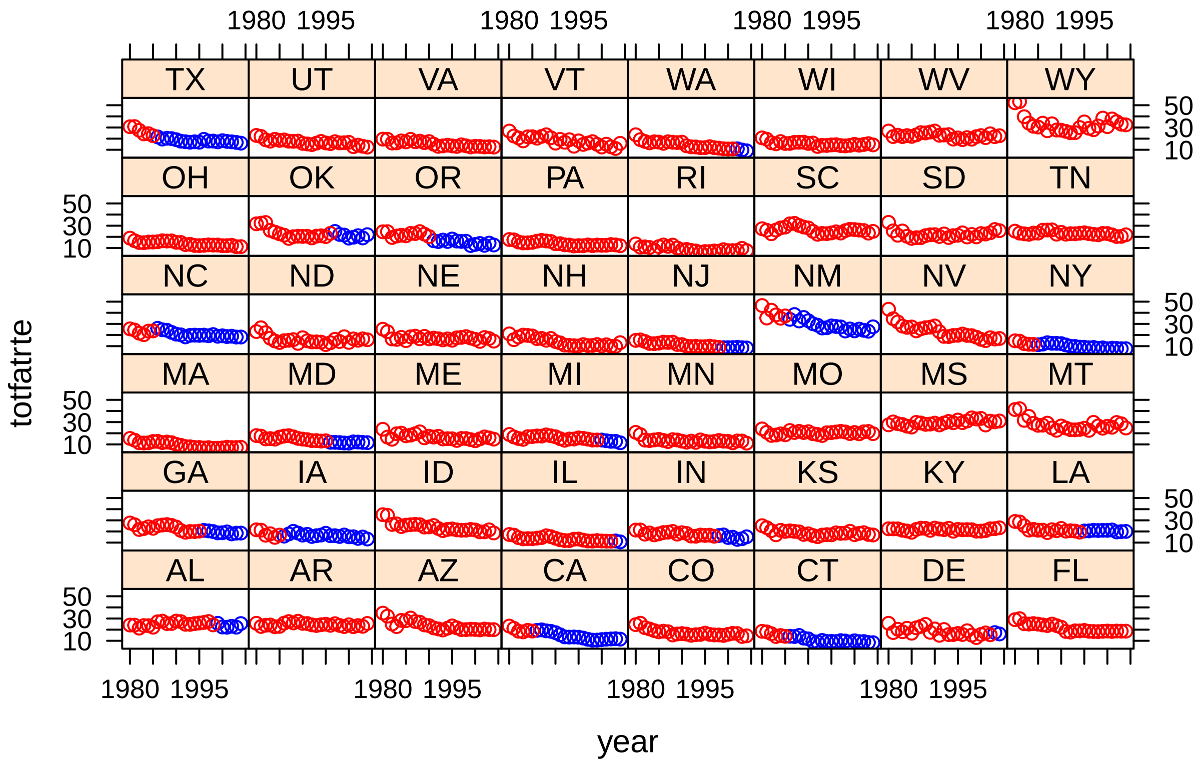



```r
# Blood alcohol limit - 8
xyplot(totfatrte~year | code, data=data_state,
       groups = bac08 < 1, col = c("blue","red"),
       main = "Figure 12: Blood alcohol limit - 8")
```




```r
# Seat belt laws - primary
xyplot(totfatrte~year | code, data=data_state,
       groups = sbprim < 1, col = c("blue","red"),
       main = "Figure 13: Seat belt laws - primary")
```

Figure 13: Seat belt laws − primary




```
# Seat belt laws - secondary
xyplot(totfatrte~year | code, data=data_state,
       groups = sbsecon < 1, col = c("blue","red"),
       main = "Figure 14: Seat belt laws - secondary")
```

Figure 14: Seat belt laws − secondary