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

Devin Robison and Lingyao Meng

U.S. traffic fatalities: 1980-2004

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.

```r
# load the RData file
load("driving.RData", f <- new.env())
# variable descriptions
# f$desc
# get the data
driving <- f$data
str(driving)
```

```
## 'data.frame':    1200 obs. of  56 variables:
##  $ year        : int  1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 ...
##  $ state       : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ sl55        : num  1 1 1 1 1 ...
##  $ sl65        : num  0 0 0 0 0 ...
##  $ sl70        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ sl75        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ slnone      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ seatbelt    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ minage      : num  18 18 18 18 18 20 21 21 21 21 ...
##  $ zerotol     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ gdl         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ bac10       : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ bac08       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ perse       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ totfat      : int  940 933 839 930 932 882 1080 1111 1024 1029 ...
##  $ nghtfat     : int  422 434 376 397 421 358 500 499 423 418 ...
##  $ wkndfat     : int  236 248 224 223 237 224 279 300 226 247 ...
##  $ totfatpvm   : num  3.2 3.35 2.81 3 2.83 ...
##  $ nghtfatpvm  : num  1.44 1.56 1.26 1.28 1.28 ...
##  $ wkndfatpvm  : num  0.803 0.89 0.75 0.719 0.72 ...
##  $ statepop    : int  3893888 3918520 3925218 3934109 3951834 3972527 3991569 4015261 402385
##  $ totfatrte   : num  24.1 24.1 21.4 23.6 23.6 ...
##  $ nghtfatrte  : num  10.84 11.08 9.58 10.09 10.65 ...
```

```
##  $ wkndfatrte  : num  6.06 6.33 5.71 5.67 6 ...
##  $ vehicmiles  : num  29.4 27.9 29.9 31 32.9 ...
##  $ unem        : num  8.8 10.7 14.4 13.7 11.1 ...
##  $ perc14_24   : num  18.9 18.7 18.4 18 17.6 ...
##  $ sl70plus    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ sbprim      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ sbsecon     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ d80         : int  1 0 0 0 0 0 0 0 0 0 ...
##  $ d81         : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ d82         : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ d83         : int  0 0 0 1 0 0 0 0 0 0 ...
##  $ d84         : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ d85         : int  0 0 0 0 0 1 0 0 0 0 ...
##  $ d86         : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ d87         : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ d88         : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ d89         : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ d90         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ d91         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ d92         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ d93         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ d94         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ d95         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ d96         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ d97         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ d98         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ d99         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ d00         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ d01         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ d02         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ d03         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ d04         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ vehicmilespc: num  7544 7108 7607 7880 8334 ...
##  - attr(*, "datalabel")= chr ""
##  - attr(*, "time.stamp")= chr "22 Jan 2013 14:09"
##  - attr(*, "formats")= chr  "%8.0g" "%8.0g" "%9.0g" "%9.0g" ...
##  - attr(*, "types")= int  252 251 254 254 254 254 254 251 254 254 ...
##  - attr(*, "val.labels")= chr  "" "" "" "" ...
##  - attr(*, "var.labels")= chr  "1980 through 2004" "48 continental states, alphabetical" "sp
##  - attr(*, "version")= int 12
```

The dataset has 1200 observations of 56 variables. The response variables are traffic fatalities. The explanatory variables include the year dummies, traffic laws enforcement dummies and some geographic and economic factors.

The response variable we are interested in is the total fatality rate and the potential explanatory variables include the year dummies, the blood alcohol concentration (BAC) limits, the seatbelt laws, the speed limit of 70 and up, the *per se* law, the graduated drivers license law, the unemployment rate, the percent population aged 14 to 24 and the vehicle miles traveled per capita.

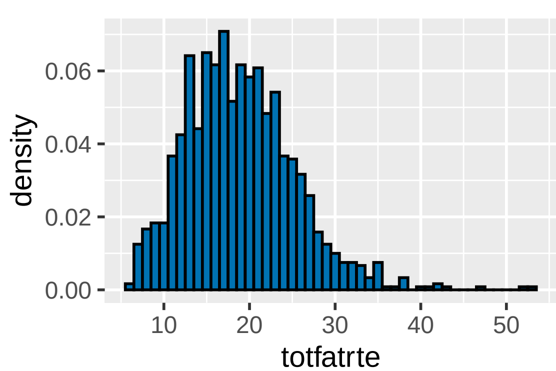
```r
summary(driving[c('totfatrte','bac08','bac10','sbprim','sbsecon','sl70plus',
                  'perse','gdl','unem','perc14_24','vehicmilespc')])
```

```
##    totfatrte         bac08            bac10            sbprim
##  Min.   : 6.20   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:14.38   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :18.43   Median :0.0000   Median :1.0000   Median :0.0000
##  Mean   :18.92   Mean   :0.2135   Mean   :0.6231   Mean   :0.1792
##  3rd Qu.:22.77   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :53.32   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##     sbsecon          sl70plus          perse             gdl
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :1.0000   Median :0.0000
##  Mean   :0.4683   Mean   :0.2068   Mean   :0.5471   Mean   :0.1741
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##       unem          perc14_24      vehicmilespc
##  Min.   : 2.200   Min.   :11.70   Min.   : 4372
##  1st Qu.: 4.500   1st Qu.:13.90   1st Qu.: 7788
##  Median : 5.600   Median :14.90   Median : 9013
##  Mean   : 5.951   Mean   :15.33   Mean   : 9129
##  3rd Qu.: 7.000   3rd Qu.:16.60   3rd Qu.:10327
##  Max.   :18.000   Max.   :20.30   Max.   :18390
```

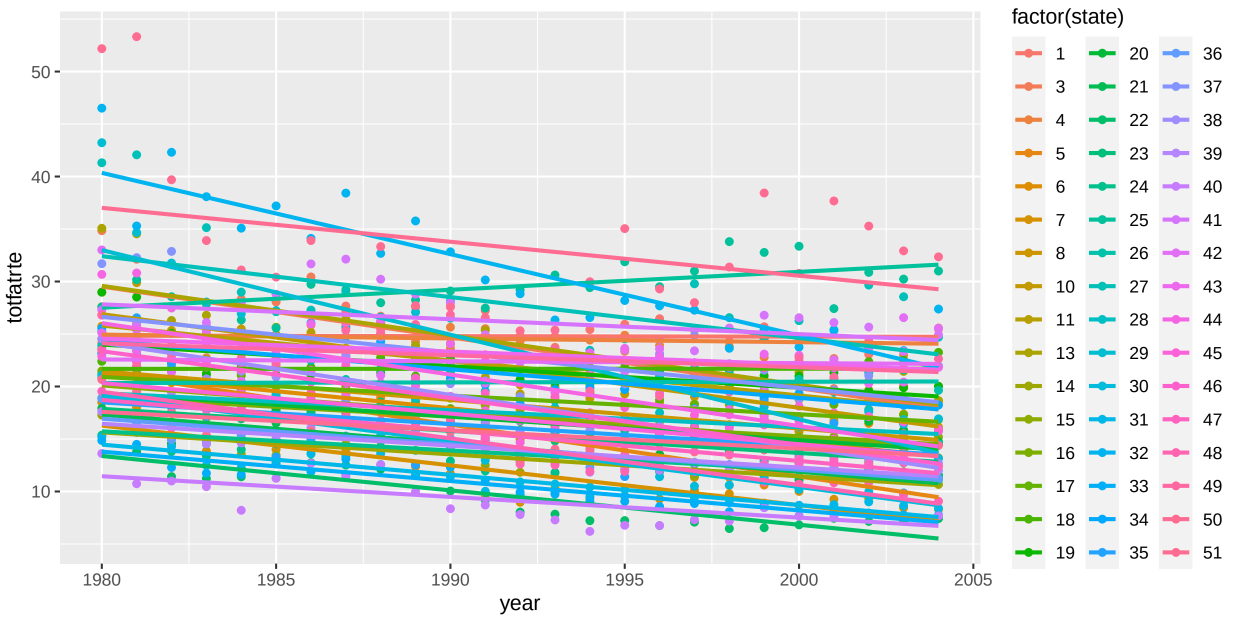
No irregular values were observed from these variables.

Univariate analysis of the response variable

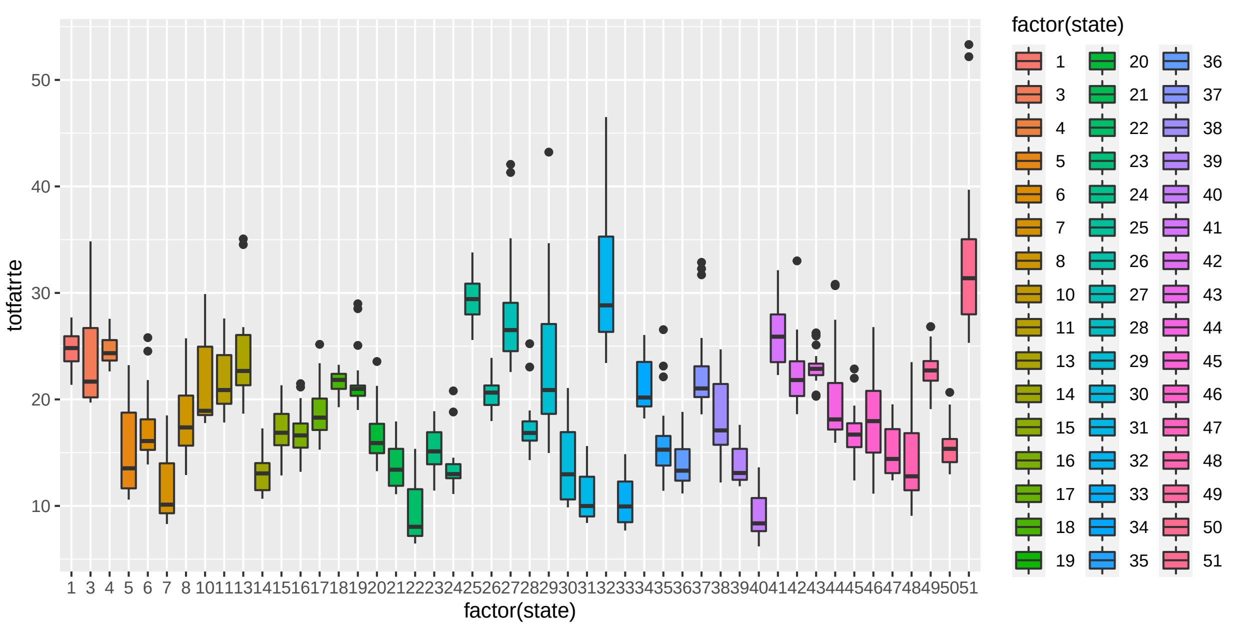
```r
ggplot(driving, aes(x = totfatrte)) +
  geom_histogram(aes(y = ..density..), binwidth = 1, fill="#0072B2", colour="black")
```



```r
ggplot(driving, aes(x = year, y = totfatrte, color = factor(state))) +
  geom_point() + geom_smooth(method=lm, se=FALSE)
```



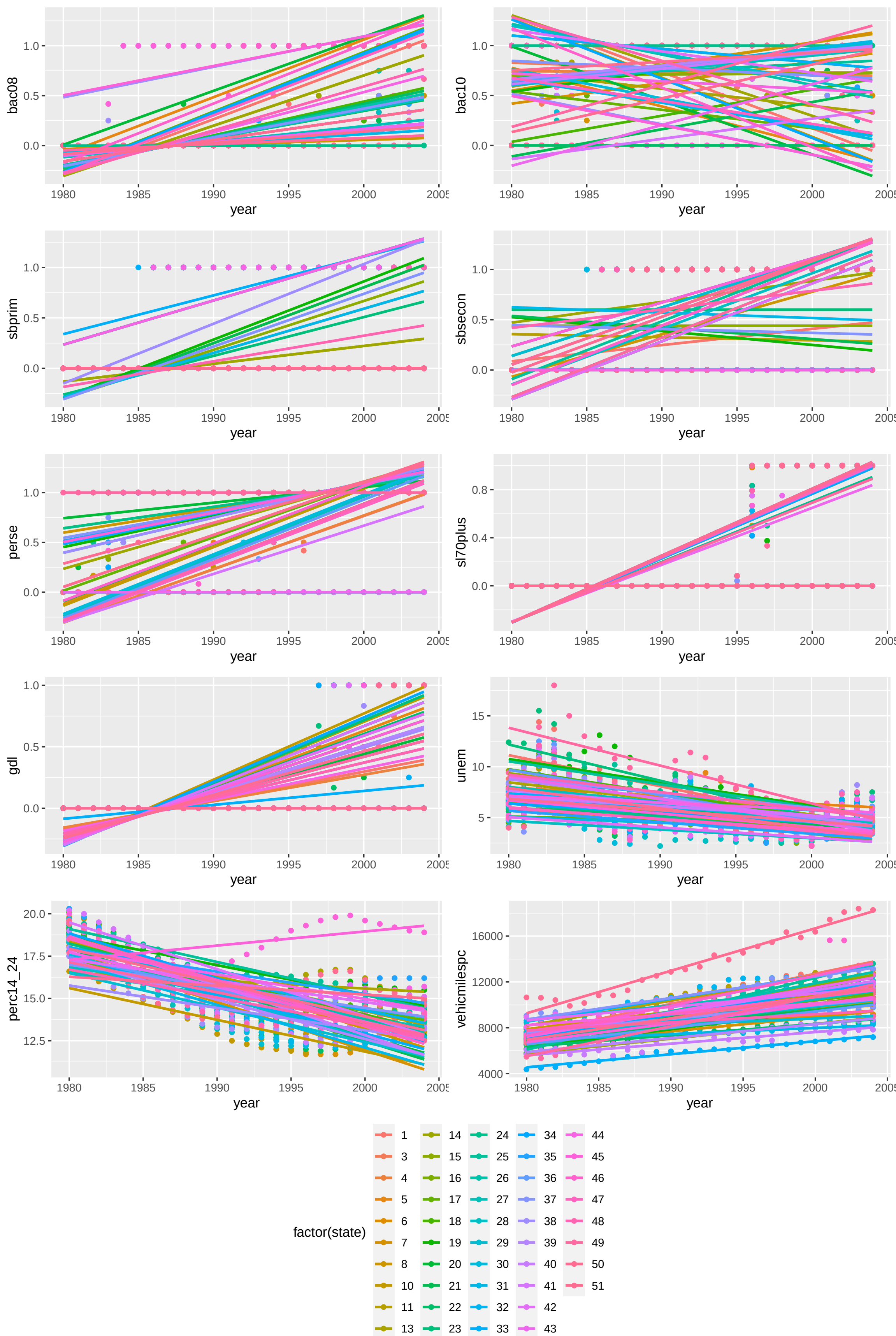

```r
ggplot(driving, aes(factor(state), totfatrte)) +
  geom_boxplot(aes(fill = factor(state)))
```



The distribution of the response variable *totfatrte* is slightly right skewed. Since the skewness is not very serious, we decided not to perform transformation on it. From the time plot grouped by state, we could see that for most states, the fatality rate trended to decrease from 1980 to 2004. From the boxplot, we observed variated data variances across states.

Univariate analysis of the explanatory variables

```
uni.bac08 <- qplot(x = year, y = bac08, data = driving, color = factor(state)) +
  geom_smooth(method=lm, se=FALSE)
uni.bac10 <- qplot(x = year, y = bac10, data = driving, color = factor(state)) +
  geom_smooth(method=lm, se=FALSE)
uni.sbprim <- qplot(x = year, y = sbprim, data = driving, color = factor(state)) +
  geom_smooth(method=lm, se=FALSE)
uni.sbsecon <- qplot(x = year, y = sbsecon, data = driving, color = factor(state)) +
  geom_smooth(method=lm, se=FALSE)
uni.perse <- qplot(x = year, y = perse, data = driving, color = factor(state)) +
  geom_smooth(method=lm, se=FALSE)
uni.sl70plus <- qplot(x = year, y = sl70plus, data = driving, color = factor(state)) +
  geom_smooth(method=lm, se=FALSE)
uni.gdl <- qplot(x = year, y = gdl, data = driving, color = factor(state)) +
  geom_smooth(method=lm, se=FALSE)
uni.unem <- qplot(x = year, y = unem, data = driving, color = factor(state)) +
  geom_smooth(method=lm, se=FALSE)
uni.perc14_24 <- qplot(x = year, y = perc14_24, data = driving, color = factor(state)) +
  geom_smooth(method=lm, se=FALSE)
uni.vehicmilespc <- qplot(x = year, y = vehicmilespc, data = driving, color = factor(state)) +
  geom_smooth(method=lm, se=FALSE)
ggarrange(uni.bac08, uni.bac10, uni.sbprim, uni.sbsecon, uni.perse, uni.sl70plus, uni.gdl, uni
          uni.perc14_24, uni.vehicmilespc, ncol=2, nrow=5, common.legend = TRUE, legend="bottom
```



bac08
bac10
sbprim
sbsecon
perse
sl70plus
gdl
unem
perc14_24
vehicmilespc
year
factor(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 30 31 32 33
34 35 36 37 38 39 40 41 42 43
44 45 46 47 48 49 50 51

From the time plots, we see the enforcement of BAC limit of 0.08% increased by time, for quite a few states. In fact, over 75% of the observations valued 0 in *bac08*. On the other hand, comparable increasing and decreasing trends were observed on the enforcement of BAC limit of 0.10%, indicating that the enforcement of two limits may not be mutually exclusive. Both variables need to be kept in the model.

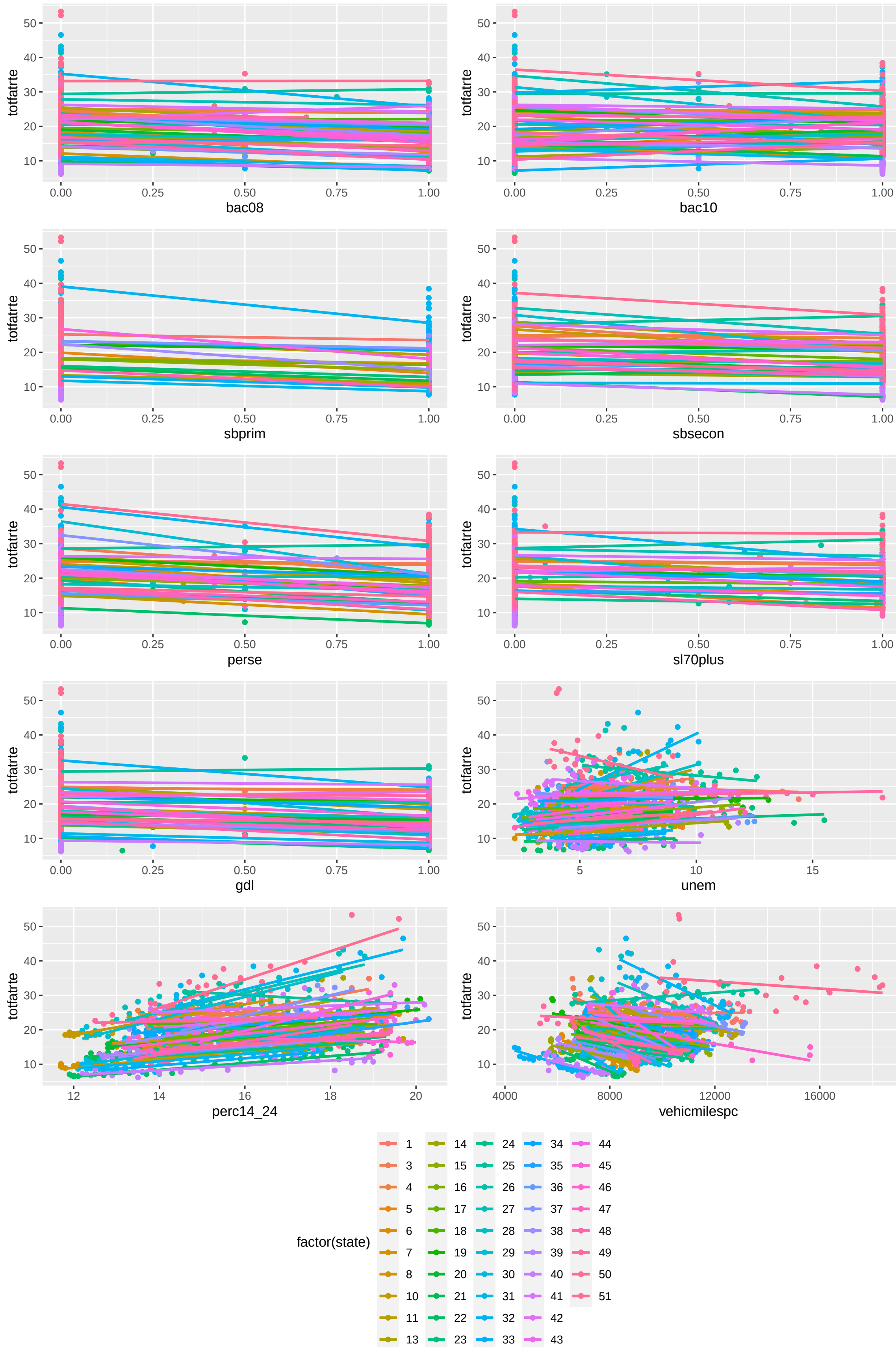
The time plot showed that the enforcement of the primary seat belt law trended to increase from 1980 to 2004 for a few states. Similar to *bac08*, over 75% of the observations valued 0 in *sbprim*. In quite a few states, we observed increase trend for the enforcement of the second seat belt law. There were some states where the trend was decrease though.

In most states, the enforcement of the "Per se" law trended to increase from 1980 to 2004. There are also some states where the law remained in effect or never in effect in the period. In a few states, the enforcement of speed limit of 70 and up trended to increase from 1980 to 2004. Some states had never enacted such high speed limit in the period. In fact, over 75% observations valued 0 in *sl70plus*. In a few states, the enforcement of the graduated drivers license law trended to increase from 1980 to 2004. Some states had never enacted the law in the period. In fact, over 75% observations valued 0 in *gdl*.

In most states, the unemployment rate and the percent population aged 14 to 24 trended to decrease from 1980 to 2004. In most states, the vehicle miles traveled per capita trended to increase from 1980 to 2004.

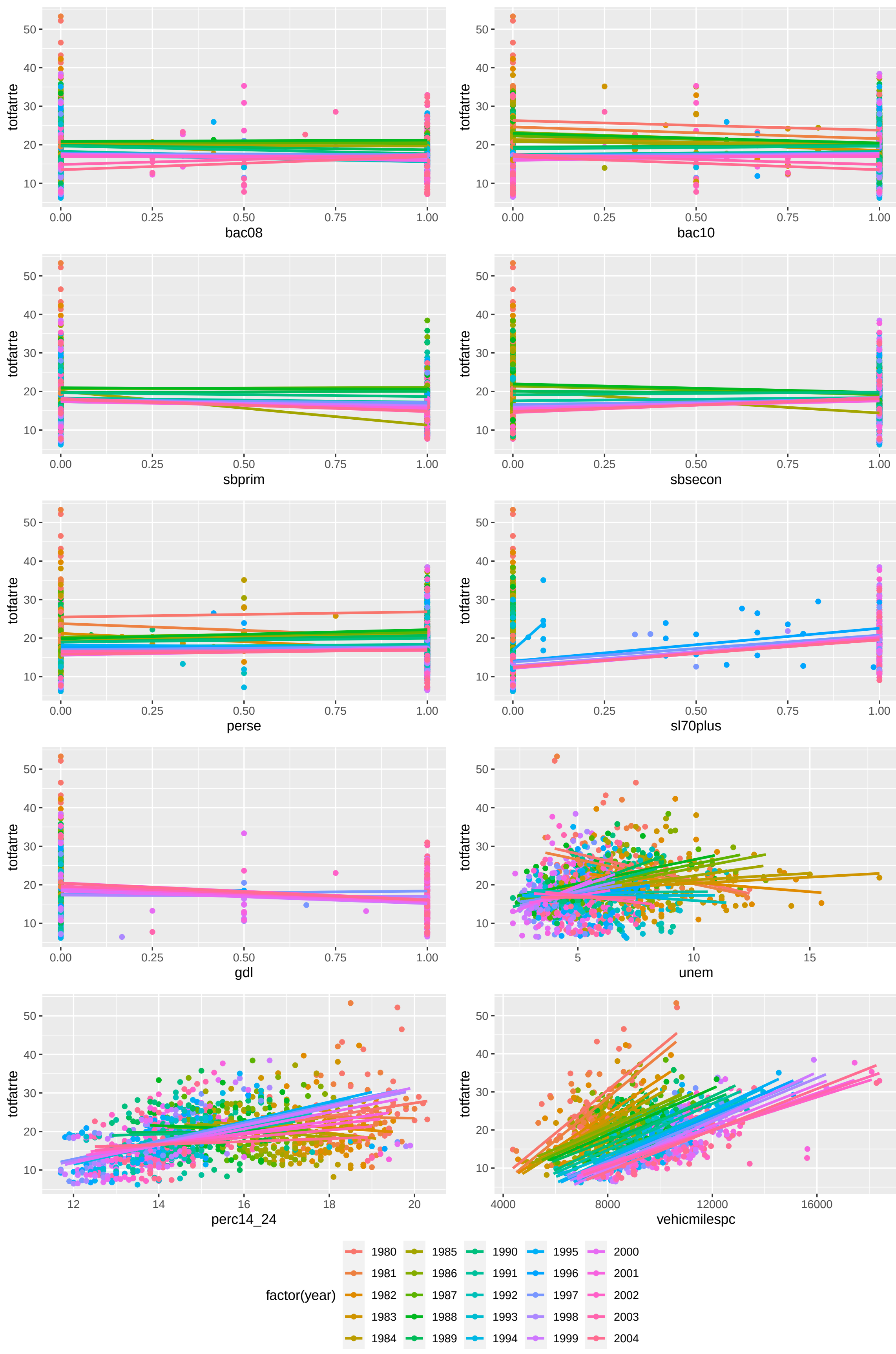
Bivariate analysis by state

```
bi.bac08.state <- qplot(x = bac08, y = totfatrte, data = driving, color = factor(state)) +
  geom_smooth(method=lm, se=FALSE)
bi.bac10.state <- qplot(x = bac10, y = totfatrte, data = driving, color = factor(state)) +
  geom_smooth(method=lm, se=FALSE)
bi.sbprim.state <- qplot(x = sbprim, y = totfatrte, data = driving, color = factor(state)) +
  geom_smooth(method=lm, se=FALSE)
bi.sbsecon.state <- qplot(x = sbsecon, y = totfatrte, data = driving, color = factor(state)) +
  geom_smooth(method=lm, se=FALSE)
bi.perse.state <- qplot(x = perse, y = totfatrte, data = driving, color = factor(state)) +
  geom_smooth(method=lm, se=FALSE)
bi.sl70plus.state <- qplot(x = sl70plus, y = totfatrte, data = driving, color = factor(state))
  geom_smooth(method=lm, se=FALSE)
bi.gdl.state <- qplot(x = gdl, y = totfatrte, data = driving, color = factor(state)) +
  geom_smooth(method=lm, se=FALSE)
bi.unem.state <- qplot(x = unem, y = totfatrte, data = driving, color = factor(state)) +
  geom_smooth(method=lm, se=FALSE)
bi.perc14_24.state <- qplot(x = perc14_24, y = totfatrte, data = driving, color = factor(state
  geom_smooth(method=lm, se=FALSE)
bi.vehicmilespc.state <- qplot(x = vehicmilespc, y = totfatrte, data = driving, color = factor
  geom_smooth(method=lm, se=FALSE)
ggarrange(bi.bac08.state, bi.bac10.state, bi.sbprim.state, bi.sbsecon.state, bi.perse.state,
          bi.sl70plus.state, bi.gdl.state, bi.unem.state, bi.perc14_24.state, bi.vehicmilespc.s
          ncol=2, nrow=5, common.legend = TRUE, legend="bottom")
```


totfatrte
bac08
bac10
sbprim
sbsecon
perse
sl70plus
gdl
unem
perc14_24
vehicmilespc
factor(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
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51

Bivariate analysis by the year

```
bi.bac08.year <- qplot(x = bac08, y = totfatrte, data = driving, color = factor(year)) +
  geom_smooth(method=lm, se=FALSE)
bi.bac10.year <- qplot(x = bac10, y = totfatrte, data = driving, color = factor(year)) +
  geom_smooth(method=lm, se=FALSE)
bi.sbprim.year <- qplot(x = sbprim, y = totfatrte, data = driving, color = factor(year)) +
  geom_smooth(method=lm, se=FALSE)
bi.sbsecon.year <- qplot(x = sbsecon, y = totfatrte, data = driving, color = factor(year)) +
  geom_smooth(method=lm, se=FALSE)
bi.perse.year <- qplot(x = perse, y = totfatrte, data = driving, color = factor(year)) +
  geom_smooth(method=lm, se=FALSE)
bi.sl70plus.year <- qplot(x = sl70plus, y = totfatrte, data = driving, color = factor(year)) +
  geom_smooth(method=lm, se=FALSE)
bi.gdl.year <- qplot(x = gdl, y = totfatrte, data = driving, color = factor(year)) +
  geom_smooth(method=lm, se=FALSE)
bi.unem.year <- qplot(x = unem, y = totfatrte, data = driving, color = factor(year)) +
  geom_smooth(method=lm, se=FALSE)
bi.perc14_24.year <- qplot(x = perc14_24, y = totfatrte, data = driving, color = factor(year))
  geom_smooth(method=lm, se=FALSE)
bi.vehicmilespc.year <- qplot(x = vehicmilespc, y = totfatrte, data = driving, color = factor(
  geom_smooth(method=lm, se=FALSE)
ggarrange(bi.bac08.year, bi.bac10.year, bi.sbprim.year, bi.sbsecon.year, bi.perse.year,
          bi.sl70plus.year, bi.gdl.year, bi.unem.year, bi.perc14_24.year, bi.vehicmilespc.year
          ncol=2, nrow=5, common.legend = TRUE, legend="bottom")
```


totfatrte
bac08
bac10
sbprim
sbsecon
perse
sl70plus
gdl
unem
perc14_24
vehicmilespc
factor(year)
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004

Within states, some negative correlation was observed between *bac08* and *totfatrte*. Within a year, the correlation is not very obvious. This suggests that for a given state, the enforcement of BAC limit of 0.08% would probably decrease the fatality rate. However, there are other effects than *bac08* in explanation of different fatality rates in a year among states.

On the other hand, the correlation between *bac10* and *totfatrte* within states is not very clear. Within a year, some negative correlation was observed.

Within states, negative correlations were observed between both the primary and the secondary seatbelt law and the fatality rate. Within a year, the correlation between *sbprim* and *totfatrte* is still negative but that between *sbsecon* and *totfatrte* is mixed.

The enforcement of the "Per se" laws was negatively correlated with the fatality rate for most states, with few exceptions. However, within a year, the correlation seems to be wealy positive.

Interestingly, we observed negative correlation within states between high speed limit (70 and up) and the fatality rate and positive correlation within a year. This suggests complicated effects of *sl70plus* on *totfatrte*.

Negative correlations were observed between the enforcement of graduated drivers license law and the fatality rate, both across states and across years.

Virtually, for most states, the unemployment rate is positively correlated with the fatality rate while negative correlations were also observed for a few states. The regression lines have variated slopes among states. Similar correlation between *unem* and *totfatrte* was observed within a year.

Virtually, for most states, the percent population aged 14 to 24 is positively correlated with the fatality rate while negative correlations were also observed for a few states. The regression lines have variated slopes among states. Similar correlation between *unem* and *totfatrte* was observed within a year.

Clearly, within a year, *vehicmilespc* and *totfatrte* is positively correlated. On the other hand, the within states correlation seems to be negative for most states. Meanwhile, we observed the cross states regression slopes get decreased by year. It suggests that the positive effect of *vehicmilespc* on *totfatrte* shrinks over time. This may explain the negative within state correlation as there are other factors dereasing 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.

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.)

The variables *bac08, bac10, perse, subprim, sbsecon, sl70plus, gdl* have value ranges from 0 to 1. In fact these are binary indicators of whether certain law was in effect in a state, in a year. The decimal values, if there is any, stand for the fraction of the year when the law was enacted within a year. The variables *perc14_24*, *unem* and *vehicmilespc* are continuous and the distributions are not severly skewed. Also, we didn't observed any obvious non-linear relationship between any explanatory variable and the response variable. Therefore, no transformation is needed for either variable.

```
lm2 <- lm(data = driving, totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89 + d90
            d91 + d92 + d93 + d94 + d95 + d96 + d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04 +
            bac08 + bac10 + perse + sbprim + sbsecon + sl70plus + gdl + perc14_24 + unem + vehi

summary(lm2)
```

```
##
## Call:
## lm(formula = totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 +
##     d87 + d88 + d89 + d90 + d91 + d92 + d93 + d94 + d95 + d96 +
##     d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04 + bac08 + bac10 +
##     perse + sbprim + sbsecon + sl70plus + gdl + perc14_24 + unem +
##     vehicmilespc, data = driving)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.9160  -2.7384  -0.2778   2.2859  21.4203
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.716e+00  2.476e+00  -1.097 0.272847
## d81          -2.175e+00  8.276e-01  -2.629 0.008686 **
## d82          -6.596e+00  8.534e-01  -7.729 2.33e-14 ***
## d83          -7.397e+00  8.690e-01  -8.512  < 2e-16 ***
## d84          -5.850e+00  8.763e-01  -6.676 3.79e-11 ***
## d85          -6.483e+00  8.948e-01  -7.245 7.82e-13 ***
## d86          -5.853e+00  9.307e-01  -6.289 4.52e-10 ***
## d87          -6.367e+00  9.670e-01  -6.585 6.87e-11 ***
## d88          -6.592e+00  1.014e+00  -6.502 1.17e-10 ***
## d89          -8.071e+00  1.053e+00  -7.667 3.68e-14 ***
## d90          -8.959e+00  1.077e+00  -8.319 2.46e-16 ***
## d91          -1.107e+01  1.101e+00 -10.052  < 2e-16 ***
## d92          -1.288e+01  1.123e+00 -11.473  < 2e-16 ***
## d93          -1.273e+01  1.136e+00 -11.204  < 2e-16 ***
## d94          -1.236e+01  1.157e+00 -10.685  < 2e-16 ***
## d95          -1.195e+01  1.184e+00 -10.098  < 2e-16 ***
## d96          -1.388e+01  1.223e+00 -11.343  < 2e-16 ***
## d97          -1.426e+01  1.250e+00 -11.408  < 2e-16 ***
## d98          -1.504e+01  1.265e+00 -11.886  < 2e-16 ***
```

```
## d99           -1.509e+01  1.284e+00 -11.750  < 2e-16 ***
## d00           -1.544e+01  1.305e+00 -11.831  < 2e-16 ***
## d01           -1.618e+01  1.334e+00 -12.131  < 2e-16 ***
## d02           -1.672e+01  1.348e+00 -12.406  < 2e-16 ***
## d03           -1.702e+01  1.359e+00 -12.521  < 2e-16 ***
## d04           -1.671e+01  1.387e+00 -12.049  < 2e-16 ***
## bac08         -2.498e+00  5.375e-01  -4.648 3.73e-06 ***
## bac10         -1.418e+00  3.963e-01  -3.577 0.000362 ***
## perse         -6.201e-01  2.982e-01  -2.079 0.037791 *
## sbprim        -7.533e-02  4.908e-01  -0.153 0.878032
## sbsecon        6.728e-02  4.293e-01   0.157 0.875492
## sl70plus       3.348e+00  4.452e-01   7.521 1.09e-13 ***
## gdl           -4.269e-01  5.269e-01  -0.810 0.417978
## perc14_24      1.416e-01  1.227e-01   1.154 0.248675
## unem           7.571e-01  7.791e-02   9.718  < 2e-16 ***
## vehicmilespc  2.925e-03  9.497e-05  30.804  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.046 on 1165 degrees of freedom
## Multiple R-squared:  0.6078, Adjusted R-squared:  0.5963
## F-statistic:  53.1 on 34 and 1165 DF,  p-value: < 2.2e-16
```

The variable *bac08* is the binary indicator of whether the blood alcohol concentration (BAC) of 0.08% was allowed in a state, in a year. The variable *bac10* is the binary indicator of whether the blood alcohol concentration of 0.10% was allowed in a state, in a year.

The coefficient of *bac08* was estimated as -2.5. It means that holding all other conditions constant, when the BAC limit of 0.08% was enforced, the total fatality rate would drop by 2.5. The coefficient of *bac10* was estimated as -1.4. It means that holding all other conditions constant, when the BAC limit of 0.10% was enforced, the total fatality rate would drop by 1.4. Clearly, the effect of imposing BAC limit of 0.08% was estimated to be larger than that of 0.10%, in decreasing the total fatality rate.

The coefficient of *perse* was estimated as -0.062 and the p-value is smaller than 0.05. There is marginal evidence to claim that the effect of *perse* on the total fatality rate is negative. On the other hand, the t-test for the coefficient of *sbprim* resulted in a quite large p-value, so there is a lack of evidence to claim that *sbprim* has effect on the total fatality rate.

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?

```r
driving.panel <- pdata.frame(driving, c('state', 'year'))

fe <- plm(data = driving.panel,
          totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89 + d90 + d91 + d92 + d
```

```r
           d94 + d95 + d96 + d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04 + bac08 + bac10 +
           perse + sbprim + sbsecon + sl70plus + gdl + perc14_24 + unem + vehicmilespc,
         model = 'within')

summary(fe)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 +
##     d87 + d88 + d89 + d90 + d91 + d92 + d93 + d94 + d95 + d96 +
##     d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04 + bac08 + bac10 +
##     perse + sbprim + sbsecon + sl70plus + gdl + perc14_24 + unem +
##     vehicmilespc, data = driving.panel, model = "within")
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -8.4273592 -1.0258600 -0.0029547  0.9572345 14.8109310
##
## Coefficients:
##                 Estimate  Std. Error  t-value  Pr(>|t|)
## d81          -1.51107133  0.41321486  -3.6569 0.0002672 ***
## d82          -3.02549578  0.44243119  -6.8383 1.316e-11 ***
## d83          -3.50360069  0.45657705  -7.6736 3.628e-14 ***
## d84          -4.25936110  0.46494255  -9.1610 < 2.2e-16 ***
## d85          -4.72679311  0.48547032  -9.7365 < 2.2e-16 ***
## d86          -3.66118539  0.51769787  -7.0721 2.686e-12 ***
## d87          -4.30578838  0.55532856  -7.7536 2.001e-14 ***
## d88          -4.76712131  0.60155650  -7.9246 5.501e-15 ***
## d89          -6.12997263  0.64019069  -9.5752 < 2.2e-16 ***
## d90          -6.22973766  0.66485076  -9.3701 < 2.2e-16 ***
## d91          -6.91714040  0.68195432 -10.1431 < 2.2e-16 ***
## d92          -7.77417239  0.70288580 -11.0604 < 2.2e-16 ***
## d93          -8.09410864  0.71594741 -11.3055 < 2.2e-16 ***
## d94          -8.50421668  0.73410866 -11.5844 < 2.2e-16 ***
## d95          -8.25540198  0.75623634 -10.9164 < 2.2e-16 ***
## d96          -8.60661913  0.79594975 -10.8130 < 2.2e-16 ***
## d97          -8.70781739  0.81975686 -10.6224 < 2.2e-16 ***
## d98          -9.34924025  0.83373487 -11.2137 < 2.2e-16 ***
## d99          -9.47489124  0.84399083 -11.2263 < 2.2e-16 ***
## d00          -9.99185979  0.85606370 -11.6719 < 2.2e-16 ***
## d01          -9.63121721  0.87255395 -11.0380 < 2.2e-16 ***
## d02          -8.90673015  0.88205263 -10.0977 < 2.2e-16 ***
## d03          -8.93650263  0.88994687 -10.0416 < 2.2e-16 ***
## d04          -9.33936116  0.91107045 -10.2510 < 2.2e-16 ***
```

```
## bac08        -1.43722116  0.39421213  -3.6458 0.0002788 ***
## bac10        -1.06266776  0.26883763  -3.9528 8.208e-05 ***
## perse        -1.15161719  0.23398721  -4.9217 9.867e-07 ***
## sbprim       -1.22739974  0.34271485  -3.5814 0.0003564 ***
## sbsecon      -0.34970784  0.25217091  -1.3868 0.1657826
## sl70plus     -0.06253283  0.26931063  -0.2322 0.8164283
## gdl          -0.41177619  0.29257391  -1.4074 0.1595790
## perc14_24     0.18712169  0.09509969   1.9676 0.0493567 *
## unem         -0.57183997  0.06057851  -9.4397 < 2.2e-16 ***
## vehicmilespc  0.00094005  0.00011104   8.4656 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    12134
## Residual Sum of Squares: 4535.3
## R-Squared:      0.62624
## Adj. R-Squared: 0.59916
## F-statistic: 55.0943 on 34 and 1118 DF, p-value: < 2.22e-16
```

```r
data.frame('Pooled.OLS' = coefficients(lm2)[c('bac08', 'bac10', 'perse', 'sbprim')],
           'Fixed.effects' = coefficients(fe)[c('bac08', 'bac10', 'perse', 'sbprim')])
```

```
##         Pooled.OLS Fixed.effects
## bac08  -2.49848306     -1.437221
## bac10  -1.41756515     -1.062668
## perse  -0.62010810     -1.151617
## sbprim -0.07533472     -1.227400
```

The estimated coefficients of *bac08, bac10, perse, and sbprim* by pooled OLS and fixed effects are listed as above. All coefficients were estimated as negative, by either model. Compared to those estimated by pooled OLS, the coefficients of *bac08* and *bac10* estimated by fixed effects got smaller in the absolute values. On the other hand, the estimated coefficients of *perse* and *sbprim* got larger. Further, *sbprim* was not statistically significant when estimated by pooled OLS, but was statistically significant when estimated by fixed effects, at 5% level.

The validity of the pooled OLS model depends on the satisfaction of the CLM assumptions of: 1. Linear in parameters; 2. Random sampling; 3. No perfect collinearity; 4. Zero conditional mean; 5. Homoskedasticity; 6. Normality.

Under the current context, CLM assumption 4, 5 and 6 can hardly be satisfied when then unobserved effects are correlated with the explanatory variables. For example, drug abuse rate could be an unoberved effect for the total fatality rate and it could be correlated with unemployment rate and the percent population aged 14 to 24.

The assumptions for the fixed effects model are as follows:

1. For each i, the model is

$$y_{it} = \beta_1 x_{it1} + ... + \beta_k x_{itk} + a_i + u_{it}, t = 1, ..., T,$$

   where the β_j are the parameters to estimate and a_i is the unobserved effect.

2. Random sampling from the cross section.
3. Each explanatory variable changes over time and no perfect collineartiy.
4. $E(u_{it}|X_i, a_i) = 0$
5. $Var(u_{it}|X_i, a_i) = VAR(u_{it}) = \sigma_u^2, for\ all\ t = 1, ..., T$
6. $Cov(u_{it}, u_{is}|X_i, a_i) = 0$
7. Conditional on X_i and a_i, the u_{it} are independent and identically distributed as $Normal(0, \sigma_u^2)$.

The fixed effects model allows for arbitrary correlation between a_i and X_i in any time period. Under the current context, we don't see serious violations to these assumptions. Therefore, the coefficients estimated by fixed effects are more reliable.

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.

```r
round(coefficients(fe)['vehicmilespc'] * 1000, 0)
```

```
## vehicmilespc
##            1
```

Holding all other conditions constant, with the number of miles driven per catipa increased by 1,000, the total fatalities per 100,000 population would increase by 1.

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?