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 40238
##  $ 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
```

From Wooldridge's rdata description (`https://rdrr.io/cran/wooldridge/man/driving.html`)

sl55: speed limit == 55 sl65: speed limit == 65 sl70: speed limit == 70 sl75: speed limit == 75 slnone: no speed limit seatbelt: =0 if none, =1 if primary, =2 if secondary minage: minimum drinking age zerotol: zero tolerance law gdl: graduated drivers license law bac10: blood alcohol limit .10 bac08: blood alcohol limit .08 perse: administrative license revocation (per se law) totfat: total traffic fatalities nghtfat: total nighttime fatalities wkndfat: total weekend fatalities totfatpvm: total fatalities per 100 million miles nghtfatpvm: nighttime fatalities per 100 million miles wkndfatpvm: weekend fatalities per 100 million miles statepop: state population totfatrte: total fatalities per 100,000 population nghtfatrte: nighttime fatalities per 100,000 population wkndfatrte: weekend accidents per 100,000 population vehicmiles: vehicle miles traveled, billions unem: unemployment rate, percent perc14_24: percent population aged 14 through 24 sl70plus: sl70 + sl75 + slnone sbprim: =1 if primary seatbelt law sbsecon: =1 if secondary seatbelt law

---

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.

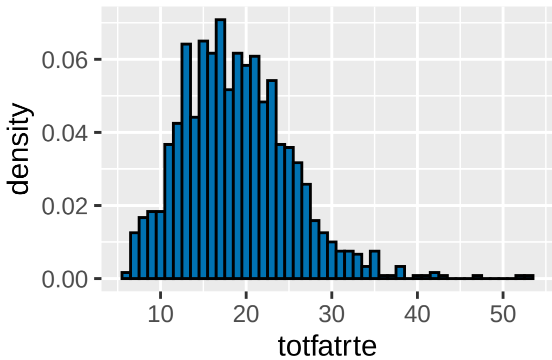
```
summary(driving[c("totfatrte", "bac08", "bac10", "sbprim", "sbsecon",
    "sl70plus", "perse", "gdl", "unem", "perc14_24", "vehicmilespc")])
```

```
##    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
```

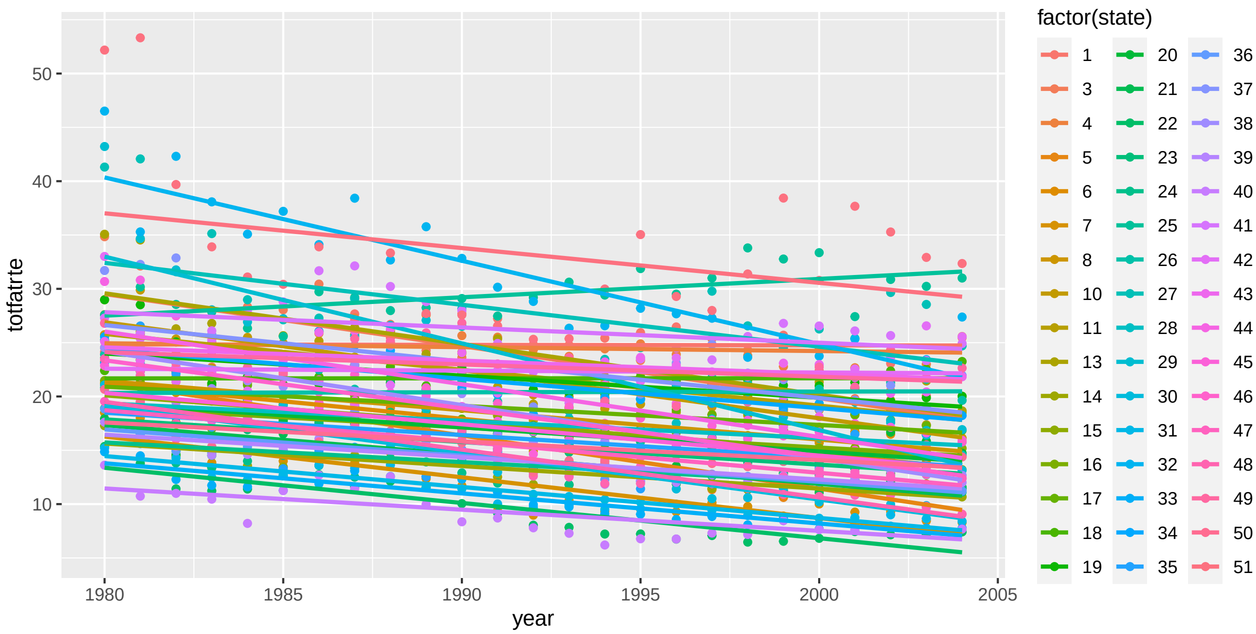
No irregular values were observed from these variables.

Univariate analysis of the response variable

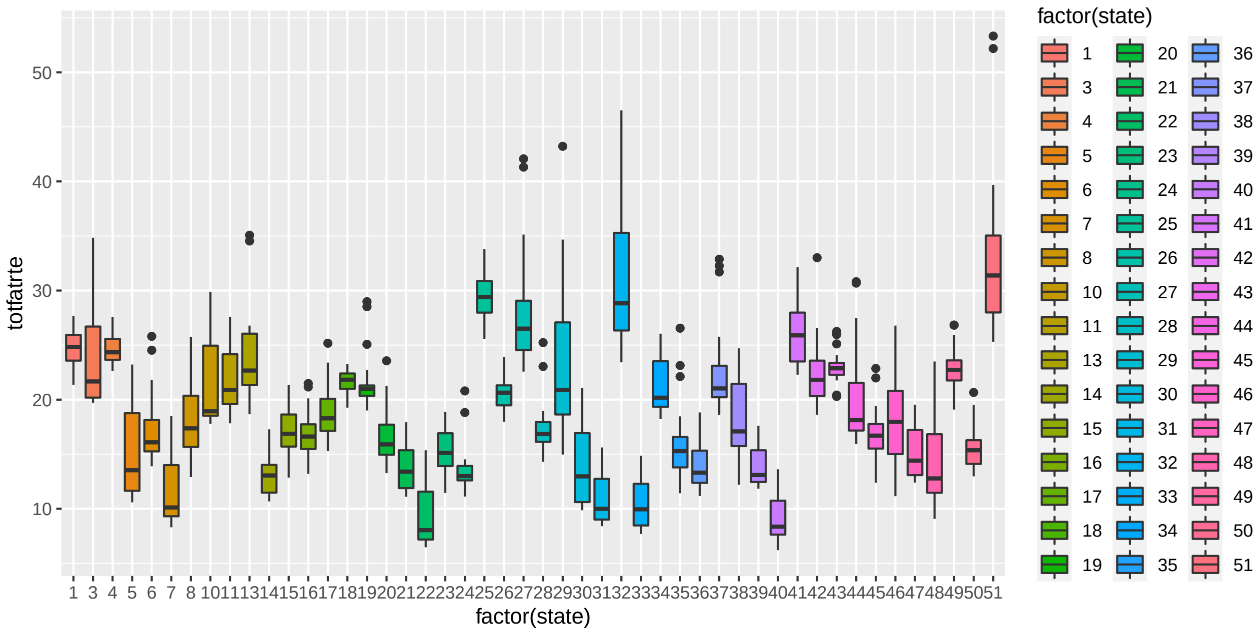
```
ggplot(driving, aes(x = totfatrte)) + geom_histogram(aes(y = ..density..),
    binwidth = 1, fill = "#0072B2", colour = "black")
```




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




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

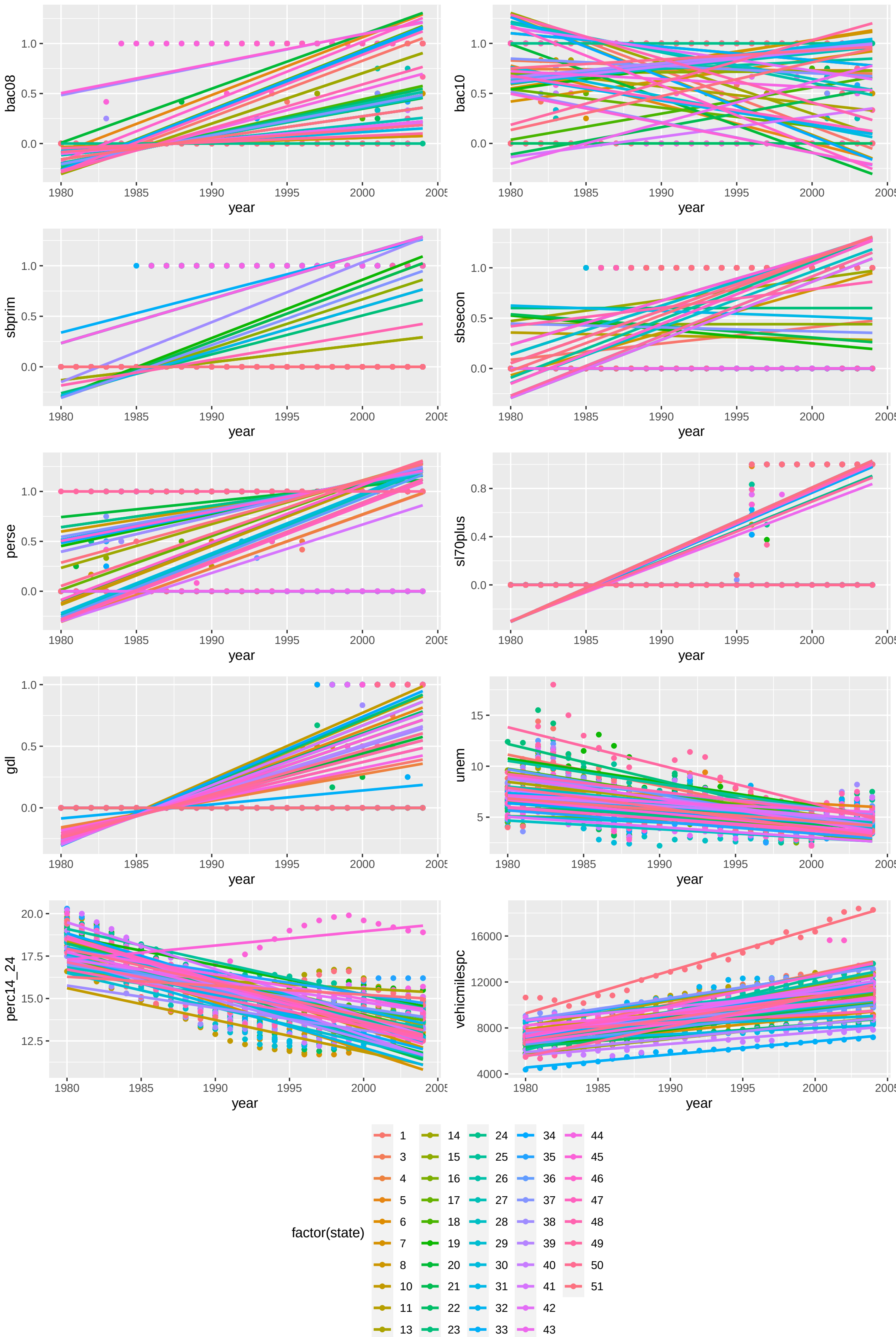


The distribution of the response variable *totfatrte* is slightly right skewed. Since the skewness is not very serious, we decided not to perform transformation on it. From the time plot grouped by state, we could see that for most states, the fatality rate trended to decrease from 1980 to 2004. From 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.unem, 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
1980
1985
1990
1995
2000
2005
0.0
0.5
1.0
0.4
0.8
5
10
15
12.5
15.0
17.5
20.0
4000
8000
12000
16000
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.

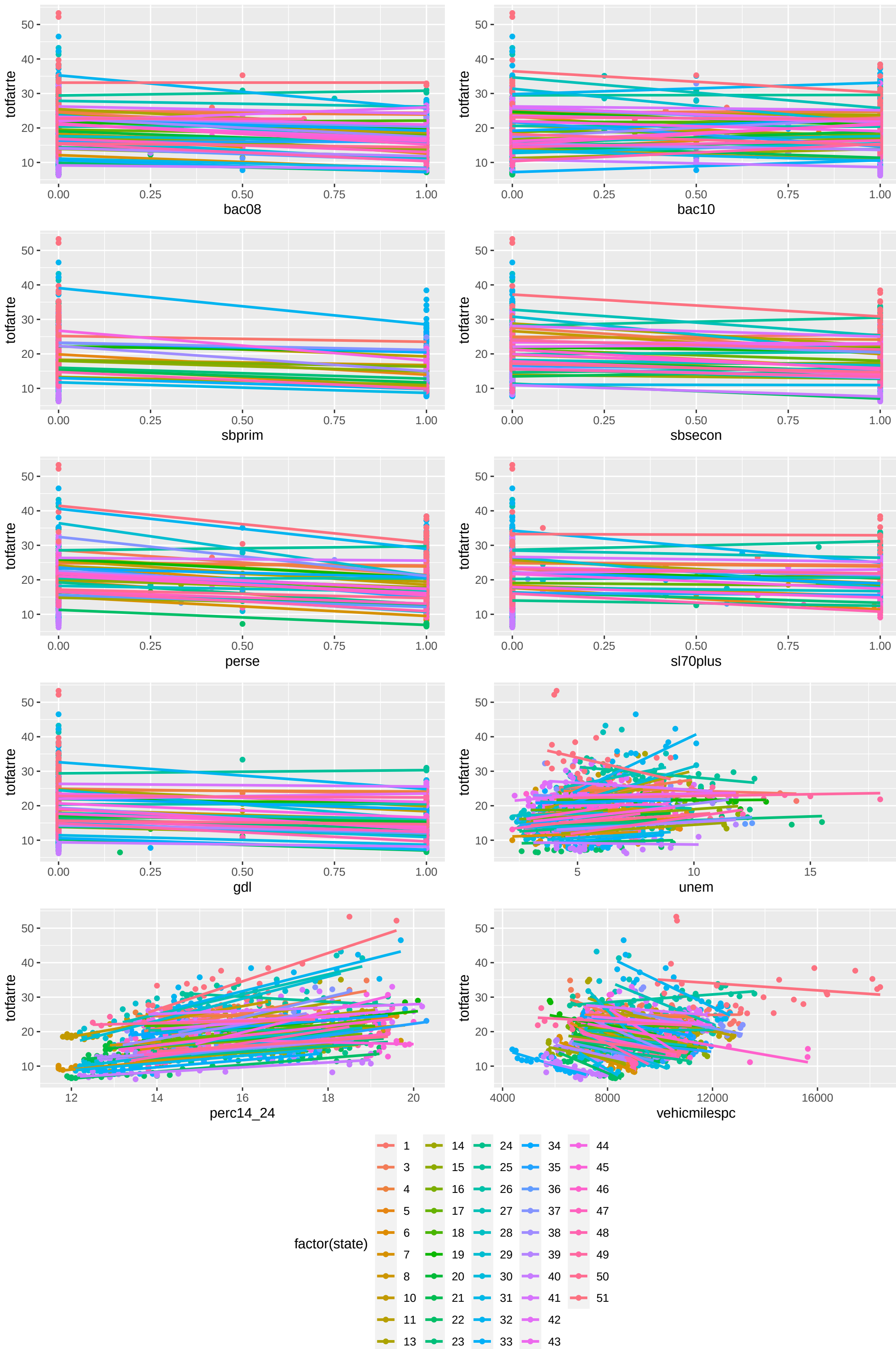
In most states, the enforcement of the "Per se" law trended to increase from 1980 to 2004. There are also some states where the law remained in effect or never in effect in the period. In a few states, the enforcement of speed limit of 70 and up trended to increase from 1980 to 2004. Some states had 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

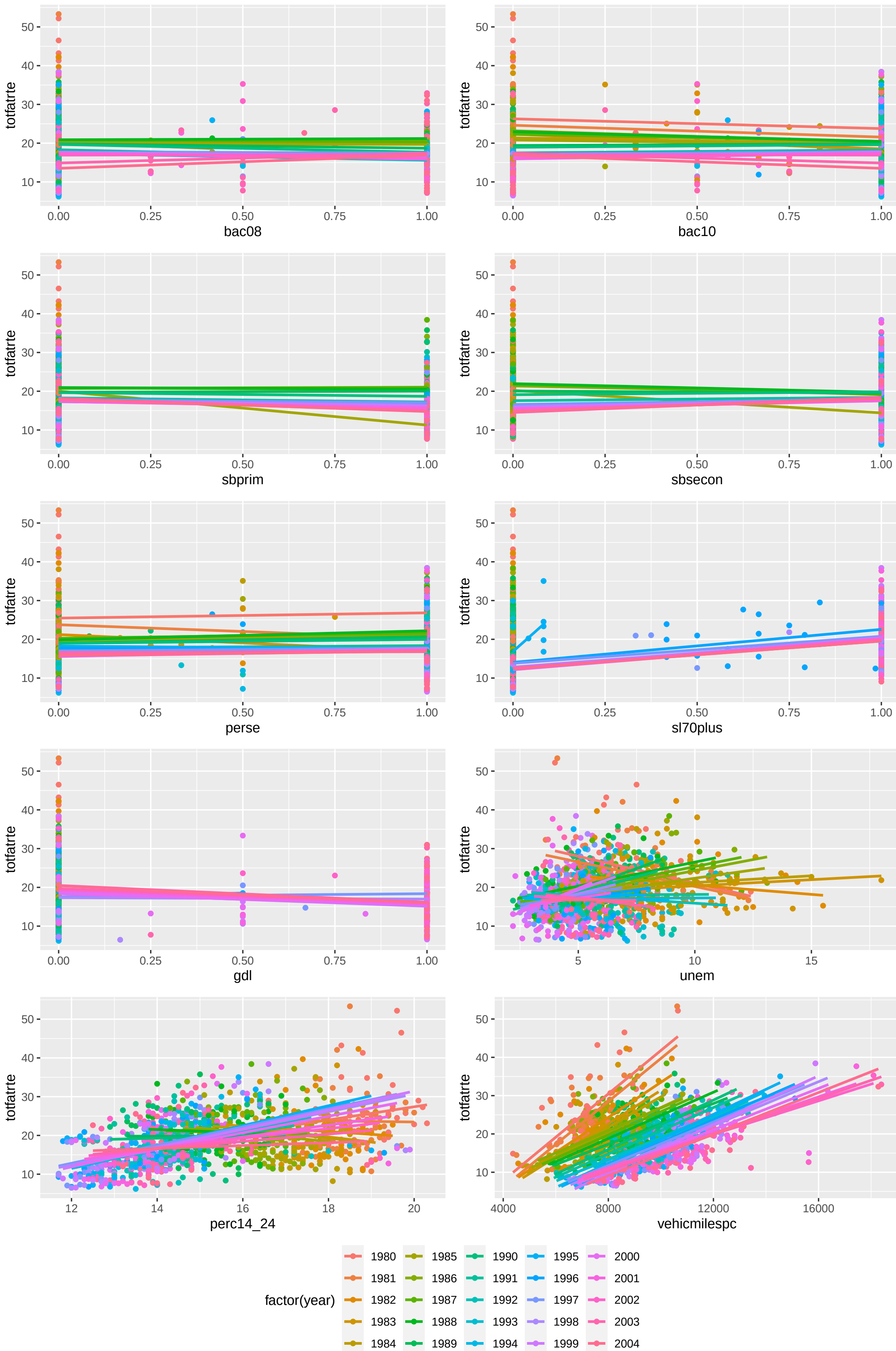
```
bi.bac08.state <- qplot(x = bac08, y = totfatrte, data = driving,
    color = factor(state)) + geom_smooth(method = lm, se = FALSE)
bi.bac10.state <- qplot(x = bac10, y = totfatrte, data = driving,
    color = factor(state)) + geom_smooth(method = lm, se = FALSE)
bi.sbprim.state <- qplot(x = sbprim, 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(state)) + 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.state, 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(year)) + 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.

```
df.traffic <- as.data.frame(f$data)
rbind(head(df.traffic, 5), tail(df.traffic, 5))
```

```
##      year state sl55 sl65 sl70 sl75 slnone seatbelt minage zerotol gdl bac10
## 1    1980     1    1    0    0    0      0        0     18       0   0   1.0
## 2    1981     1    1    0    0    0      0        0     18       0   0   1.0
## 3    1982     1    1    0    0    0      0        0     18       0   0   1.0
## 4    1983     1    1    0    0    0      0        0     18       0   0   1.0
## 5    1984     1    1    0    0    0      0        0     18       0   0   1.0
## 1196 2000    51    0    0    0    1      0        2     21       1   0   1.0
## 1197 2001    51    0    0    0    1      0        2     21       1   0   1.0
## 1198 2002    51    0    0    0    1      0        2     21       1   0   0.5
## 1199 2003    51    0    0    0    1      0        2     21       1   0   0.0
## 1200 2004    51    0    0    0    1      0        2     21       1   0   0.0
##      bac08 perse totfat nghtfat wkndfat totfatpvm nghtfatpvm wkndfatpvm
```

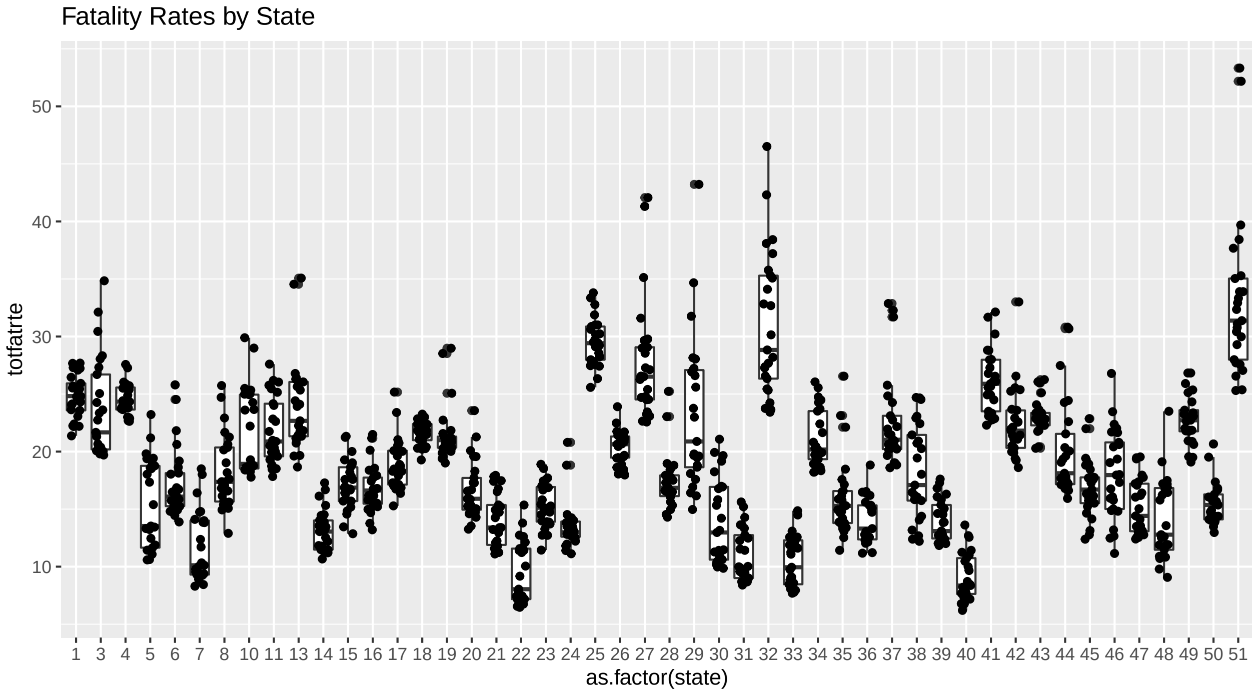
```
## 1       0.0      0     940      422      236       3.20       1.437       0.803
## 2       0.0      0     933      434      248       3.35       1.558       0.890
## 3       0.0      0     839      376      224       2.81       1.259       0.750
## 4       0.0      0     930      397      223       3.00       1.281       0.719
## 5       0.0      0     932      421      237       2.83       1.278       0.720
## 1196    0.0      1     152       59       37       1.88       0.730       0.458
## 1197    0.0      1     186       76       49       2.16       0.883       0.569
## 1198    0.5      1     176       60       29       1.95       0.665       0.321
## 1199    1.0      1     165       62       32       1.79       0.673       0.347
## 1200    1.0      1     164       67       31       1.77       0.723       0.335
##        statepop totfatrte nghtfatrte wkndfatrte vehicmiles unem perc14_24
## 1       3893888     24.14      10.84   6.060000  29.375000  8.8      18.9
## 2       3918520     24.07      11.08   6.330000  27.851999 10.7      18.7
## 3       3925218     21.37       9.58   5.710000  29.857651 14.4      18.4
## 4       3934109     23.64      10.09   5.670000  31.000000 13.7      18.0
## 5       3951834     23.58      10.65   6.000000  32.932858 11.1      17.6
## 1196     493782     30.78      11.95   7.490000   8.085110  3.9      16.1
## 1197     493754     37.67      15.39   9.920000   8.611111  3.9      15.5
## 1198     498830     35.28      12.03   5.809999   9.025640  4.2      15.3
## 1199     501242     32.92      12.37   6.380000   9.217880  4.4      15.1
## 1200     507000     32.35      13.21   6.110000   9.266000  3.7      14.9
##        sl70plus sbprim sbsecon d80 d81 d82 d83 d84 d85 d86 d87 d88 d89 d90 d91
## 1             0      0       0   1   0   0   0   0   0   0   0   0   0   0   0
## 2             0      0       0   0   1   0   0   0   0   0   0   0   0   0   0
## 3             0      0       0   0   0   1   0   0   0   0   0   0   0   0   0
## 4             0      0       0   0   0   0   1   0   0   0   0   0   0   0   0
## 5             0      0       0   0   0   0   0   1   0   0   0   0   0   0   0
## 1196          1      0       1   0   0   0   0   0   0   0   0   0   0   0   0
## 1197          1      0       1   0   0   0   0   0   0   0   0   0   0   0   0
## 1198          1      0       1   0   0   0   0   0   0   0   0   0   0   0   0
## 1199          1      0       1   0   0   0   0   0   0   0   0   0   0   0   0
## 1200          1      0       1   0   0   0   0   0   0   0   0   0   0   0   0
##        d92 d93 d94 d95 d96 d97 d98 d99 d00 d01 d02 d03 d04 vehicmilespc
## 1        0   0   0   0   0   0   0   0   0   0   0   0   0     7543.874
## 2        0   0   0   0   0   0   0   0   0   0   0   0   0     7107.785
## 3        0   0   0   0   0   0   0   0   0   0   0   0   0     7606.622
## 4        0   0   0   0   0   0   0   0   0   0   0   0   0     7879.802
## 5        0   0   0   0   0   0   0   0   0   0   0   0   0     8333.562
## 1196     0   0   0   0   0   0   0   0   1   0   0   0   0    16373.844
## 1197     0   0   0   0   0   0   0   0   0   1   0   0   0    17440.082
## 1198     0   0   0   0   0   0   0   0   0   0   1   0   0    18093.619
## 1199     0   0   0   0   0   0   0   0   0   0   0   1   0    18390.080
## 1200     0   0   0   0   0   0   0   0   0   0   0   0   1    18276.135
```

```r
# Create a restricted data set that doesn't include year
# dummies for analysis with reduced spam.
df.traffic.restricted <- df.traffic %>%
    select(!matches("d[0-9]{2}"))
```

```r
# This dumps too much for inclusion in knit
# describe(df.traffic.restricted)

df.traffic.means <- df.traffic.restricted %>%
    group_by(year) %>%
    summarise_at(vars(totfatrte, vehicmilespc, nghtfatrte, wkndfatrte),
        list(mean = mean))
df.traffic.means
```

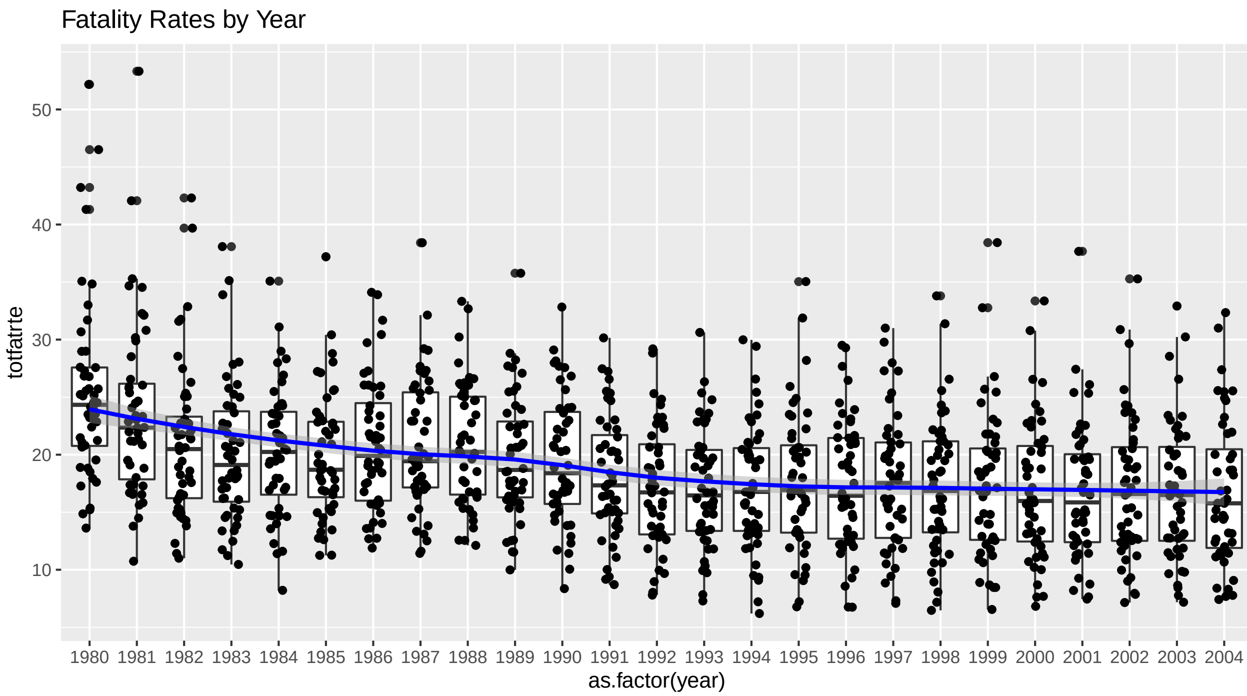
```
## # A tibble: 25 x 5
##     year totfatrte_mean vehicmilespc_mean nghtfatrte_mean wkndfatrte_mean
##    * <int>          <dbl>             <dbl>           <dbl>           <dbl>
##  1  1980           25.5             7121.           13.8             7.50
##  2  1981           23.7             7141.           12.7             6.91
##  3  1982           20.9             7259.           10.9             5.94
##  4  1983           20.2             7418.            9.98            5.36
##  5  1984           20.3             7622.            9.74            5.14
##  6  1985           19.9             7785.            9.49            4.98
##  7  1986           20.8             7966.           10.1             5.31
##  8  1987           20.8             8344.            9.78            5.20
##  9  1988           20.9             8692.            9.76            5.30
## 10  1989           19.8             8927.            9.01            4.75
## # ... with 15 more rows
```

```r
# Our ultimate question revolves around the relationship
# between traffic laws and total fotality rate so lets take a
# look at how the fatality rates have changed over time to
# get a feel for years where we might have had changes.

# Fatality rate's by state
ggplot(df.traffic) + aes(as.factor(state), totfatrte) + geom_boxplot() +
    geom_jitter(width = 0.2) + ggtitle("Fatality Rates by State")
```



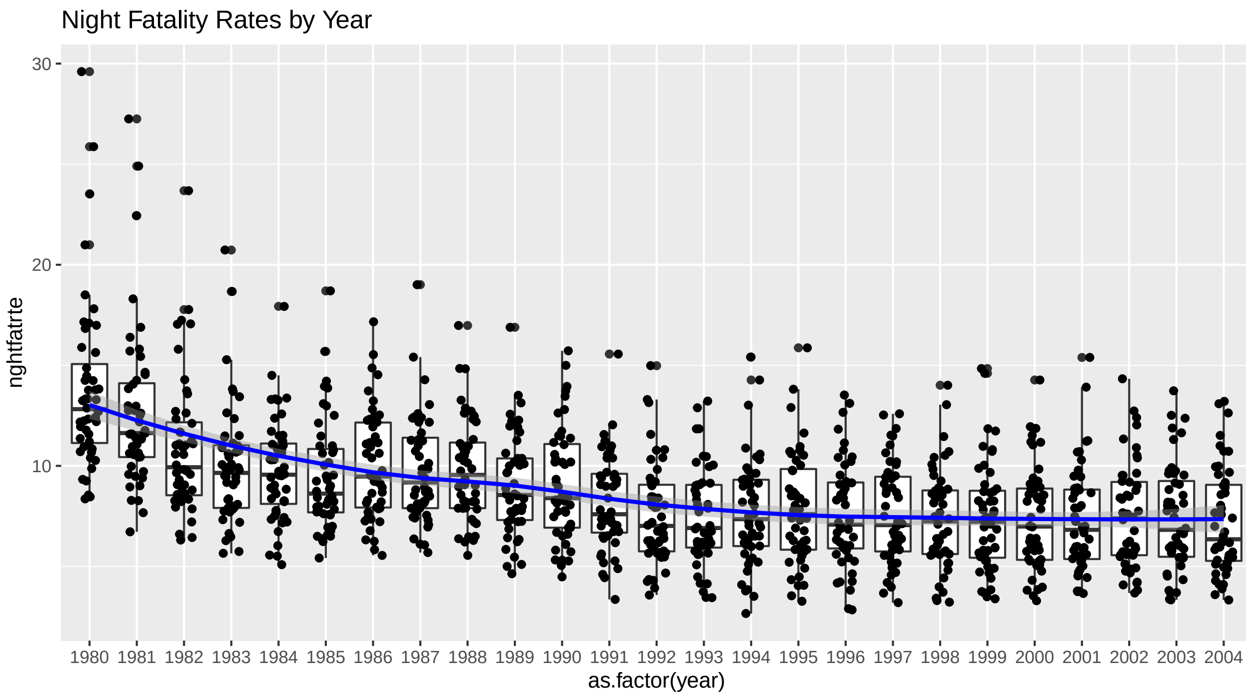

```r
# Pooled fatality rate's by year
ggplot(df.traffic) + aes(as.factor(year), totfatrte) + geom_boxplot() +
    geom_jitter(width = 0.2) + ggtitle("Fatality Rates by Year") +
    geom_smooth(data = df.traffic.means, aes(x = as.factor(year),
        y = totfatrte_mean, group = 2), color = "blue")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



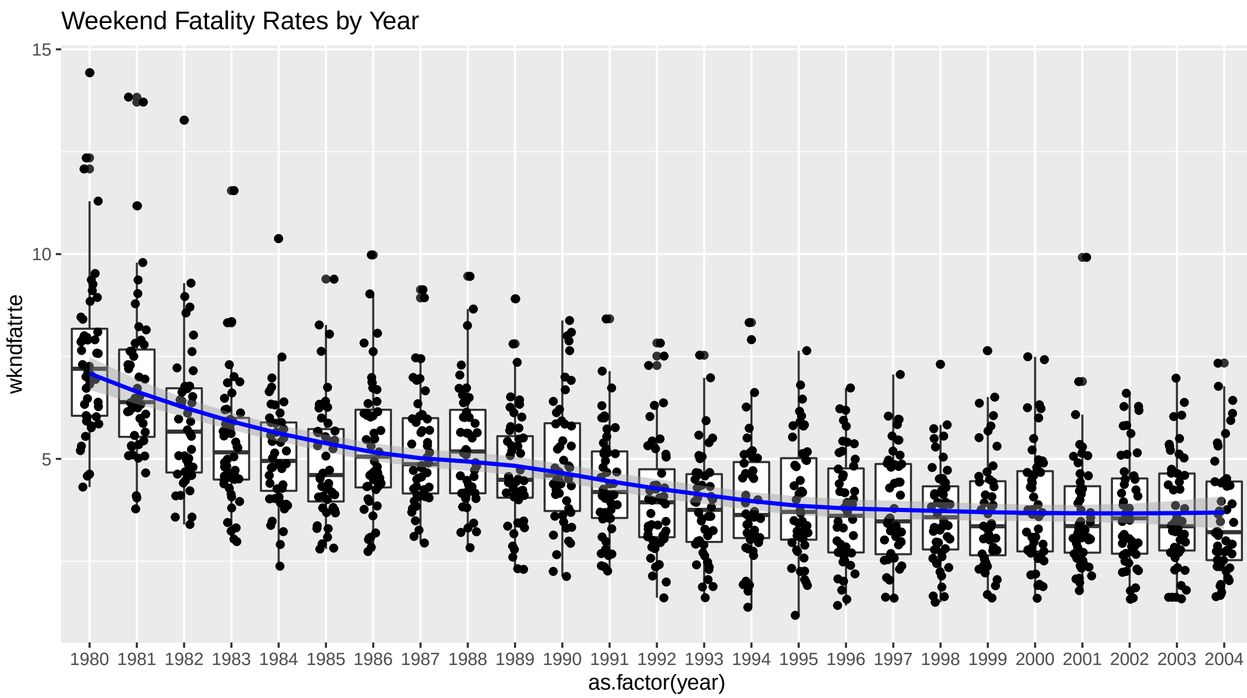
```r
ggplot(df.traffic) + aes(as.factor(year), nghtfatrte) + geom_boxplot() +
    geom_jitter(width = 0.2) + ggtitle("Night Fatality Rates by Year") +
    geom_smooth(data = df.traffic.means, aes(x = as.factor(year),
        y = nghtfatrte_mean, group = 2), color = "blue")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



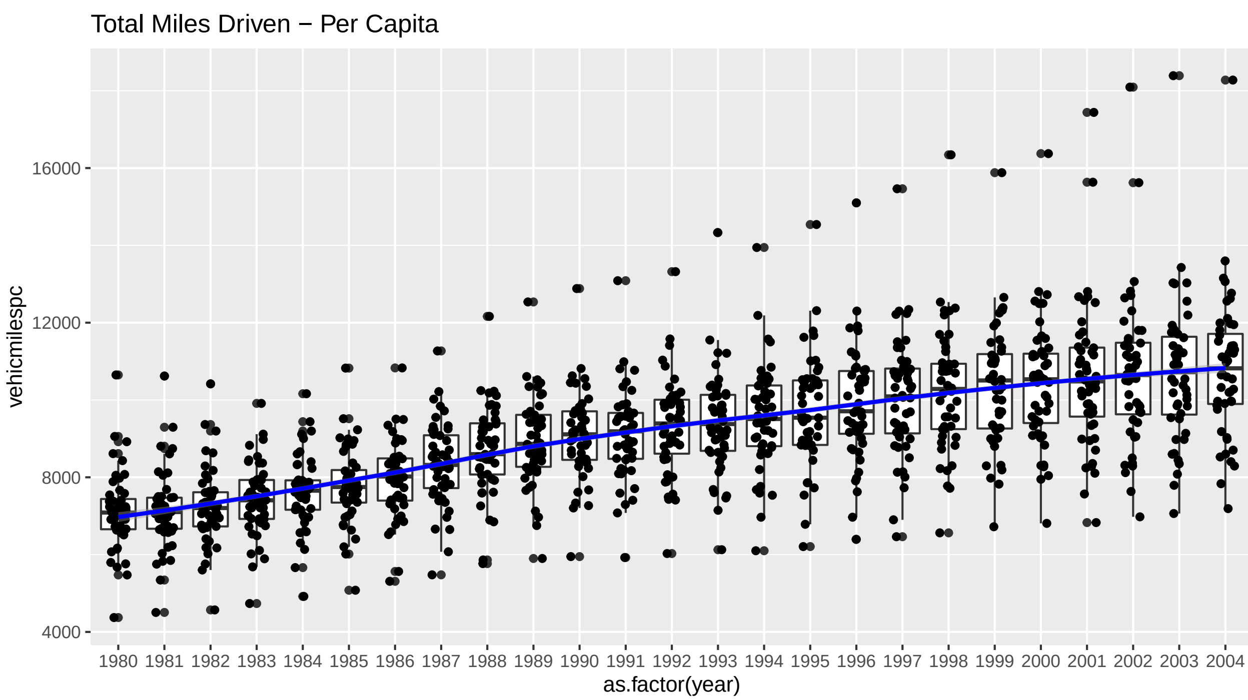

```r
ggplot(df.traffic) + aes(as.factor(year), wkndfatrte) + geom_boxplot() +
    geom_jitter(width = 0.2) + ggtitle("Weekend Fatality Rates by Year") +
    geom_smooth(data = df.traffic.means, aes(x = as.factor(year),
        y = wkndfatrte_mean, group = 2), color = "blue")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



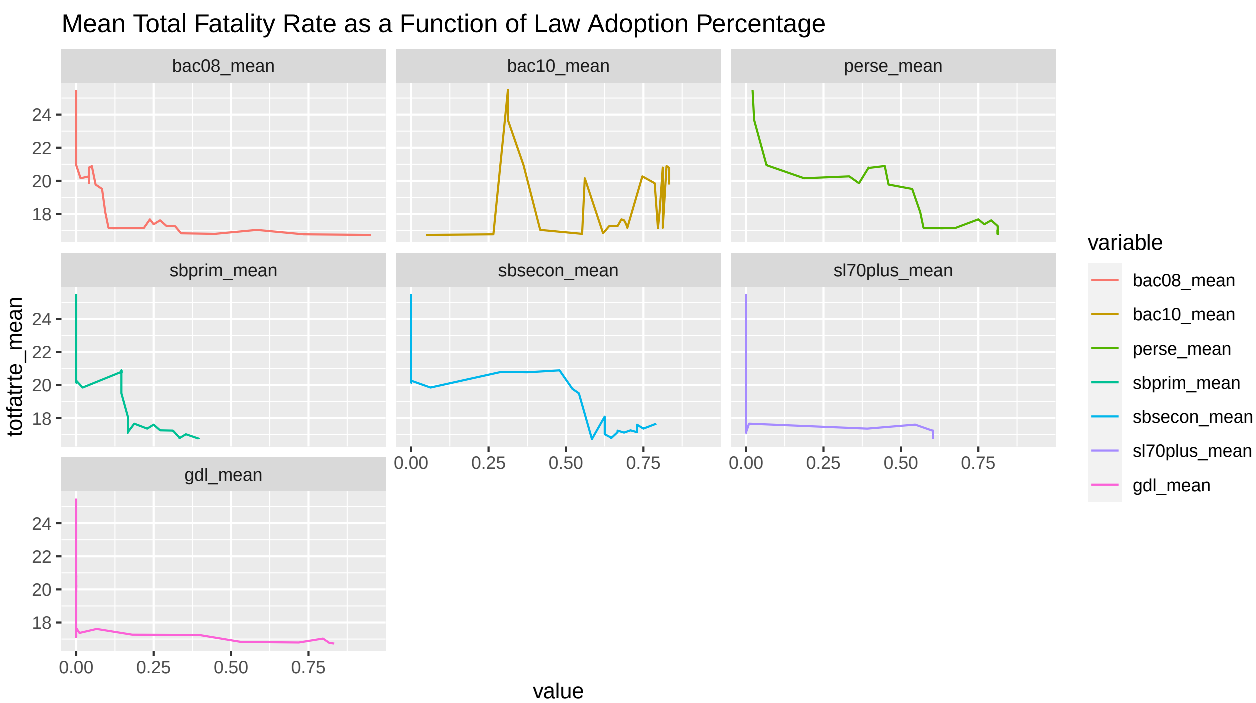

```r
ggplot(df.traffic) + aes(as.factor(year), vehicmilespc) + geom_boxplot() +
    geom_jitter(width = 0.2) + ggtitle("Total Miles Driven - Per Capita") +
    geom_smooth(data = df.traffic.means, aes(x = as.factor(year),
        y = vehicmilespc_mean, group = 1), color = "blue")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

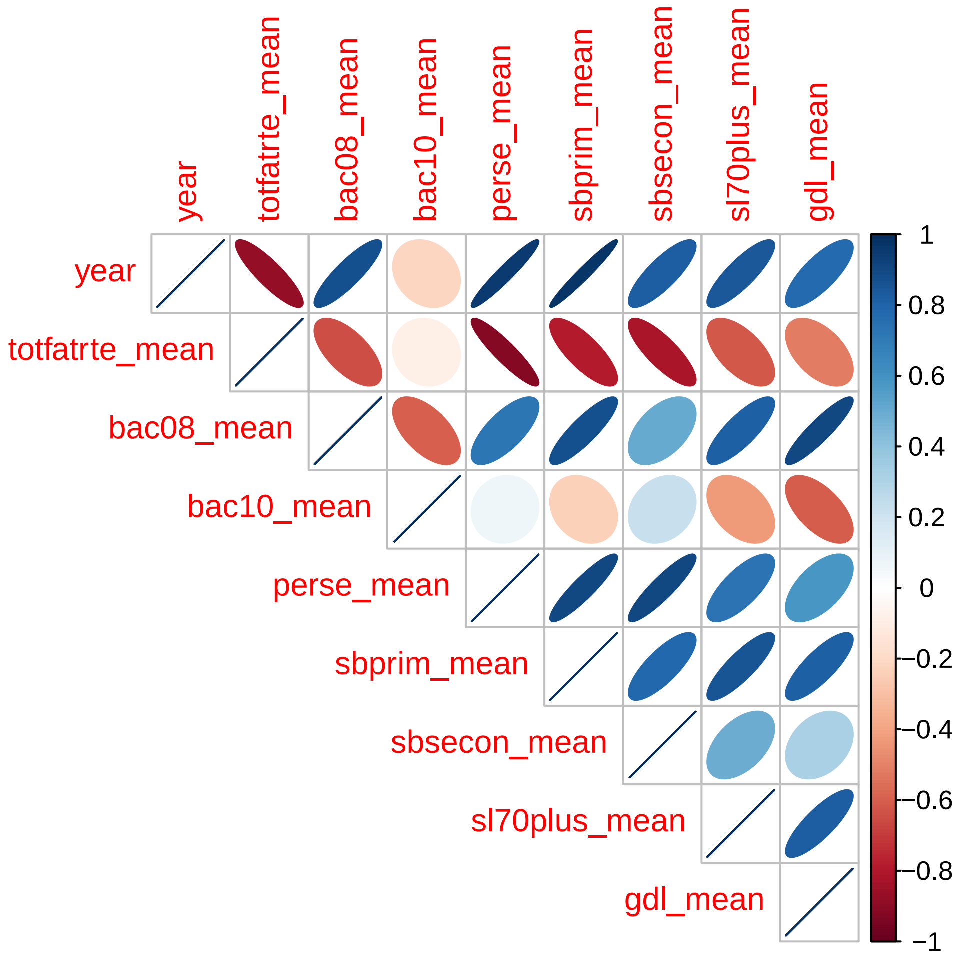



```r
# Take a look at overall correlations with specific laws.
# Since law variables are binary, we can think of their
```

```r
# average as something like an adoption rate across states in
# a given year.
df.traffic.law.means <- df.traffic.restricted %>%
    group_by(year) %>%
    mutate(perc14_24 = perc14_24/100) %>%
    mutate(unem = unem/100) %>%
    summarise_at(vars(totfatrte, bac08, bac10, perse, sbprim,
        sbsecon, sl70plus, gdl), list(mean = mean))

df.traffic.adoption <- as.data.frame(df.traffic.law.means)

df.traffic.law.melt <- melt(df.traffic.adoption, id.vars = c("year",
    "totfatrte_mean"))

ggplot(df.traffic.law.melt) + aes(value, totfatrte_mean, col = variable) +
    geom_line() + facet_wrap(~variable) + labs(title = "Mean Total Fatality Rate as a Function
```




```r
cormat <- cor(df.traffic.law.means)
corrplot(cormat, method = "ellipse", type = "upper")
```



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.

```
# Average totfatrate by year
df.traffic.means
```

```
## # A tibble: 25 x 5
##     year totfatrte_mean vehicmilespc_mean nghtfatrte_mean wkndfatrte_mean
##   * <int>          <dbl>             <dbl>           <dbl>           <dbl>
##  1  1980           25.5             7121.           13.8             7.50
##  2  1981           23.7             7141.           12.7             6.91
##  3  1982           20.9             7259.           10.9             5.94
##  4  1983           20.2             7418.            9.98            5.36
##  5  1984           20.3             7622.            9.74            5.14
##  6  1985           19.9             7785.            9.49            4.98
##  7  1986           20.8             7966.           10.1             5.31
##  8  1987           20.8             8344.            9.78            5.20
##  9  1988           20.9             8692.            9.76            5.30
## 10  1989           19.8             8927.            9.01            4.75
```

```
## # ... with 15 more rows
```

```r
# Extract dummies
df.traffic.dummies <- df.traffic %>%
    filter(year > 1980) %>%
    select(!matches("d80")) %>%
    select(matches("totfatrte") | matches("d[0-9]{2}"))
# df.traffic.dummies

lm.yr.dummie.fit <- lm(df.traffic.dummies)
# lm.yr.dummie.fit

summary(lm.yr.dummie.fit)
```

```
##
## Call:
## lm(formula = df.traffic.dummies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9302  -4.3399  -0.6952   3.7907  29.6498
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.72896    0.85289  19.614  < 2e-16 ***
## d81          6.94125    1.20617   5.755 1.12e-08 ***
## d82          4.21354    1.20617   3.493 0.000496 ***
## d83          3.42396    1.20617   2.839 0.004611 **
## d84          3.53854    1.20617   2.934 0.003417 **
## d85          3.12250    1.20617   2.589 0.009756 **
## d86          4.07146    1.20617   3.376 0.000762 ***
## d87          4.04583    1.20617   3.354 0.000822 ***
## d88          4.16271    1.20617   3.451 0.000579 ***
## d89          3.04333    1.20617   2.523 0.011768 *
## d90          2.77625    1.20617   2.302 0.021533 *
## d91          1.36583    1.20617   1.132 0.257717
## d92          0.42896    1.20617   0.356 0.722178
## d93          0.39875    1.20617   0.331 0.741013
## d94          0.42625    1.20617   0.353 0.723860
## d95          0.93958    1.20617   0.779 0.436153
## d96          0.64042    1.20617   0.531 0.595556
## d97          0.88167    1.20617   0.731 0.464952
## d98          0.53646    1.20617   0.445 0.656576
## d99          0.52146    1.20617   0.432 0.665586
## d00          0.09667    1.20617   0.080 0.936137
## d01          0.06375    1.20617   0.053 0.957858
## d02          0.30062    1.20617   0.249 0.803220
## d03          0.03458    1.20617   0.029 0.977131
```

```
## d04                NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.909 on 1128 degrees of freedom
## Multiple R-squared:  0.0931, Adjusted R-squared:  0.0746
## F-statistic: 5.034 on 23 and 1128 DF,  p-value: 1.521e-13
```

This model estimates the linear intercept for the pooled fatality rates across states, with a different intercept for each year. Quite interestingly, we see the dummy variables for 1981-1990 are at least marginally significant, an observation that appears correlated with the leveling off of the mean fatality rate across states, observed in the `Fatality rates by year` graph generated in question (1).

Given this model and the referenced graph from (1), it does appear that if we interpret a drop in totfatrte as 'driving becoming safer', then there does seem to be a general trend in that direction; additionally and without asserting any specific cause, it does appear that something(s) occurred in the 1981 to 1990 time frame that is strongly correlated with a decrease in fatality rates. Additionally, we can also observe from the `Total Miles Driven - Per Capita` graph, that the total miles being driven, per person, have gone up steadily over the same time period which, implicitly, would create more opportunities for fatal incidents to occur.

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.

```r
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 + vehicmilespc)

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 + d93 +
    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
```

```
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.

Lets start by estimating a random effects model for the same dataset.

```r
rnd_e <- plm(data = driving.panel, 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, model = "random")

summary(rnd_e)
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## 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 = "random")
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Effects:
##                 var std.dev share
## idiosyncratic 4.057   2.014 0.328
## individual    8.294   2.880 0.672
## theta: 0.8615
##
## Residuals:
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -8.25582 -1.15221 -0.15787  0.93086 16.45691
##
## Coefficients:
##                 Estimate  Std. Error  z-value  Pr(>|z|)
## (Intercept)   1.7149e+01  2.0964e+00   8.1801 2.835e-16 ***
## d81          -1.5489e+00  4.2830e-01  -3.6164 0.0002988 ***
## d82          -3.2433e+00  4.5772e-01  -7.0858 1.383e-12 ***
## d83          -3.7447e+00  4.7212e-01  -7.9318 2.161e-15 ***
## d84          -4.3729e+00  4.8064e-01  -9.0981 < 2.2e-16 ***
## d85          -4.8609e+00  5.0136e-01  -9.6954 < 2.2e-16 ***
## d86          -3.8295e+00  5.3416e-01  -7.1693 7.539e-13 ***
## d87          -4.5014e+00  5.7213e-01  -7.8678 3.610e-15 ***
## d88          -4.9819e+00  6.1887e-01  -8.0500 8.279e-16 ***
## d89          -6.3713e+00  6.5797e-01  -9.6833 < 2.2e-16 ***
## d90          -6.5357e+00  6.8279e-01  -9.5720 < 2.2e-16 ***
## d91          -7.3027e+00  7.0030e-01 -10.4279 < 2.2e-16 ***
## d92          -8.2390e+00  7.2126e-01 -11.4230 < 2.2e-16 ***
## d93          -8.5418e+00  7.3449e-01 -11.6296 < 2.2e-16 ***
## d94          -8.9183e+00  7.5297e-01 -11.8442 < 2.2e-16 ***
```

```
## d95          -8.6769e+00  7.7541e-01 -11.1902 < 2.2e-16 ***
## d96          -9.0969e+00  8.1573e-01 -11.1518 < 2.2e-16 ***
## d97          -9.2203e+00  8.3984e-01 -10.9786 < 2.2e-16 ***
## d98          -9.8922e+00  8.5380e-01 -11.5860 < 2.2e-16 ***
## d99          -1.0032e+01  8.6426e-01 -11.6071 < 2.2e-16 ***
## d00          -1.0549e+01  8.7667e-01 -12.0330 < 2.2e-16 ***
## d01          -1.0274e+01  8.9336e-01 -11.5000 < 2.2e-16 ***
## d02          -9.6376e+00  9.0278e-01 -10.6755 < 2.2e-16 ***
## d03          -9.6828e+00  9.1090e-01 -10.6300 < 2.2e-16 ***
## d04          -1.0054e+01  9.3254e-01 -10.7816 < 2.2e-16 ***
## bac08        -1.5693e+00  4.0384e-01  -3.8860 0.0001019 ***
## bac10        -1.1380e+00  2.7604e-01  -4.1227 3.744e-05 ***
## perse        -1.0933e+00  2.3885e-01  -4.5772 4.712e-06 ***
## sbprim       -1.1761e+00  3.5144e-01  -3.3465 0.0008184 ***
## sbsecon      -3.4758e-01  2.6024e-01  -1.3356 0.1816862
## sl70plus      2.9969e-02  2.7772e-01   0.1079 0.9140655
## gdl          -3.8524e-01  3.0249e-01  -1.2736 0.2028095
## perc14_24     1.9695e-01  9.7213e-02   2.0259 0.0427722 *
## unem         -4.9238e-01  6.1839e-02  -7.9622 1.690e-15 ***
## vehicmilespc  1.1744e-03  1.0983e-04  10.6933 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    12834
## Residual Sum of Squares: 5078.6
## R-Squared:      0.60429
## Adj. R-Squared: 0.59274
## Chisq: 1779.05 on 34 DF, p-value: < 2.22e-16
```

If we examine the predicted coeficients for our fixed effect model, compared with a random effects model, in a similar fashion to what was done in (4), we obtain the following:

```r
data.frame(Fixed.effects = coefficients(fe)[c("bac08", "bac10",
    "perse", "sbprim")], Random.effects = coefficients(rnd_e)[c("bac08",
    "bac10", "perse", "sbprim")])
```

```
##        Fixed.effects Random.effects
## bac08      -1.437221      -1.569317
## bac10      -1.062668      -1.138042
## perse      -1.151617      -1.093281
## sbprim     -1.227400      -1.176078
```

This indicates that the overall predictions provided by the random effects model are in close agreement with the fixed effects model. Given that this is the case, if we want to decide which model is more appropriate, we should consider the assumptions associated with each model. A major downside to our random effects model is that we must we willing to make the very strong assumption that our unobserved errors are uncorrelated with any of our explanatory variables: an extremely dubious assumption in this case. Given that both our tests produce similar model coefficients, selecting the fixed effects model seems like the correct choice on the surface.

A more structured approach, as suggested by Wooldridge, is to utilize the Hausman test for panel data, under which a rejection of the null hypothesis indicates that there is not sufficient evidence to believe that our unobserved effects are uncorrelated with our explanatory variables.

```r
phtest(fe, rnd_e)
```

```
##
##  Hausman Test
##
## data:  totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89 +  ...
## chisq = 148.69, df = 34, p-value = 2.727e-16
## alternative hypothesis: one model is inconsistent
```

The results of the Hausman test are quite stark, with an extremely small p-value, indicating that we should strongly prefer the fixed effects model.

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?

In terms of the fixed effect model, if we have serial correlation and/or heteroskedasticity in our idiosyncratic error, it generally means that we have failed to include some important time varying term, and in so doing we have violated the strictly exogenous assumption required for the model. In the case where these assumptions are violated, then standard errors and test statistics are likely invalid.

Woolridge asserts (pg.421) that it is possible to correct for serial correlation and heteroskedacity when N >> T, and N >> 1; but, then goes on to indicate that such an approach is outside the scope of the current text, so we will assume for our situation that this is the case.

In the simplifed case, where we have heteroskedacity in our idiosyncratic error without any serial correlation, we can utilize previously discussed techniques for generating robust standard errors to obtain appropriate statistics.