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. Conduct a very thorough EDA.

```
# load data
load("driving.RData", f <- new.env())
driving <- f$data
str(driving)
```

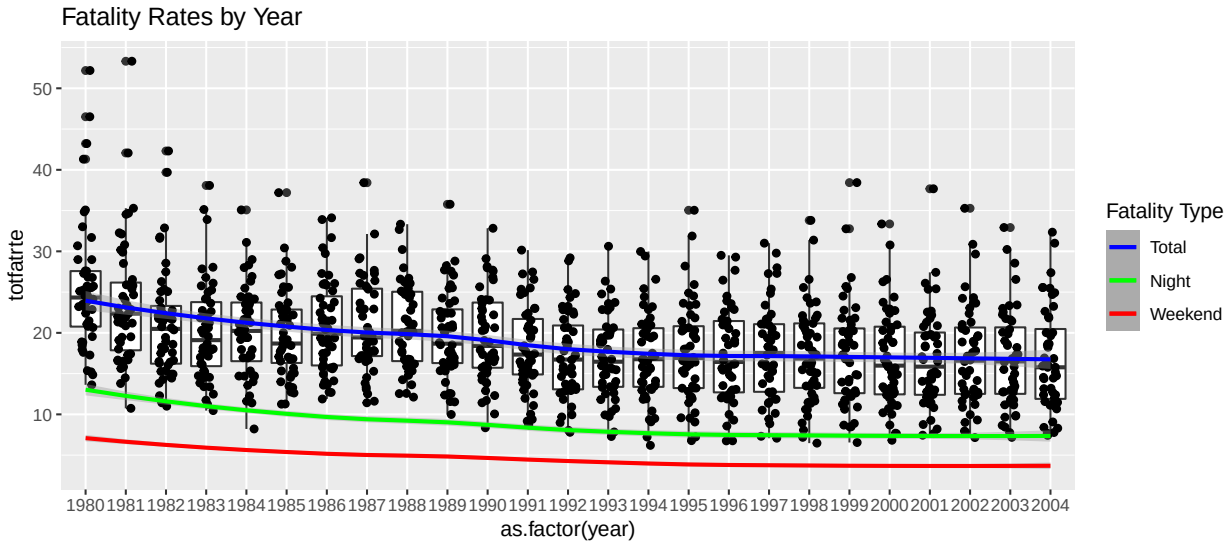
```
## 'data.frame':    1200 obs. of  56 variables:
##  $ year        : int  1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 ...
##  $ state       : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ sl55        : num  1 1 1 1 1 ...
##  $ sl65        : num  0 0 0 0 0 ...
##  $ sl70        : num  0 0 0 0 0 0 0 0 0 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
```

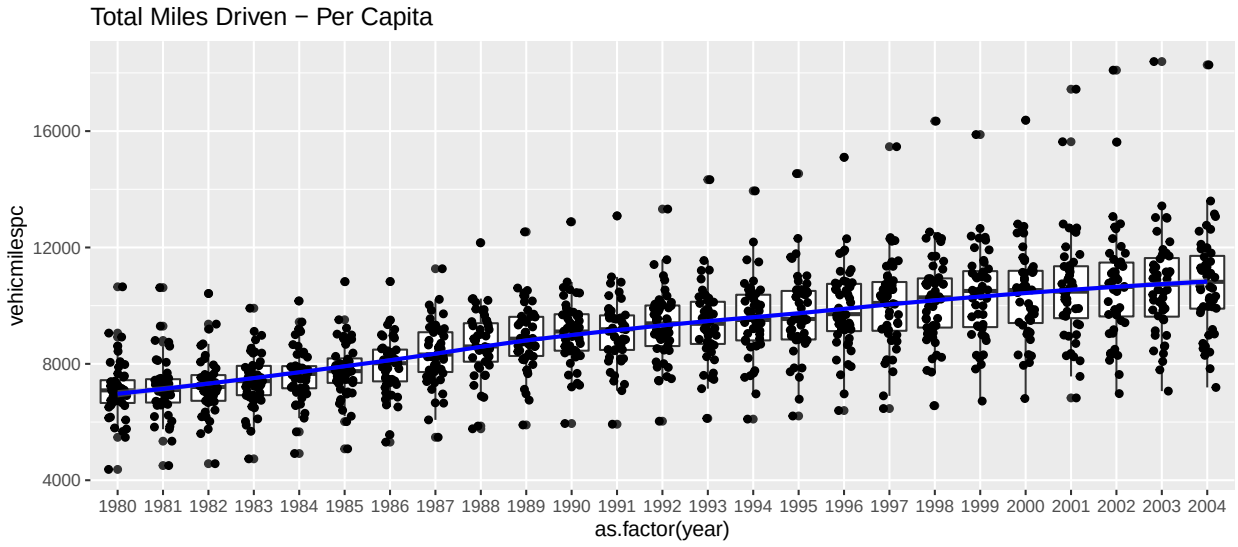
For a full variable description, see Wooldridge: `https://rdrr.io/cran/wooldridge/man/driving.html`
The dataset has 1200 observations of 56 variables, including the traffic fatalities, the year dummies, traffic laws enforcement dummies and the related geographic and economic factors. The response variable 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
# Create a restricted data set that doesn't include year dummies for analysis with reduced spa
driving.restricted <- driving %>% select(!matches("d[0-9]{2}"))
driving.means <- driving.restricted %>% group_by(year) %>%
  summarise_at(vars(totfatrte, vehicmilespc, nghtfatrte, wkndfatrte), list(mean=mean))
# Pooled fatality rate's by year
ggplot(driving) + aes(as.factor(year), totfatrte) + geom_boxplot() +
```

```
  geom_jitter(width = 0.2) + ggtitle('Fatality Rates by Year') +
  geom_smooth(data=driving.means, aes(x=as.factor(year), y=totfatrte_mean, group=2, color='blue
  geom_smooth(data=driving.means, aes(x=as.factor(year), y=nghtfatrte_mean, group=2, color='gre
  geom_smooth(data=driving.means, aes(x=as.factor(year), y=wkndfatrte_mean, group=2, color='red
  scale_colour_manual(name = 'Fatality Type',
         values =c('blue'='blue','green'='green', 'red'='red'), labels = c('Total', 'Night', 'W
```




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

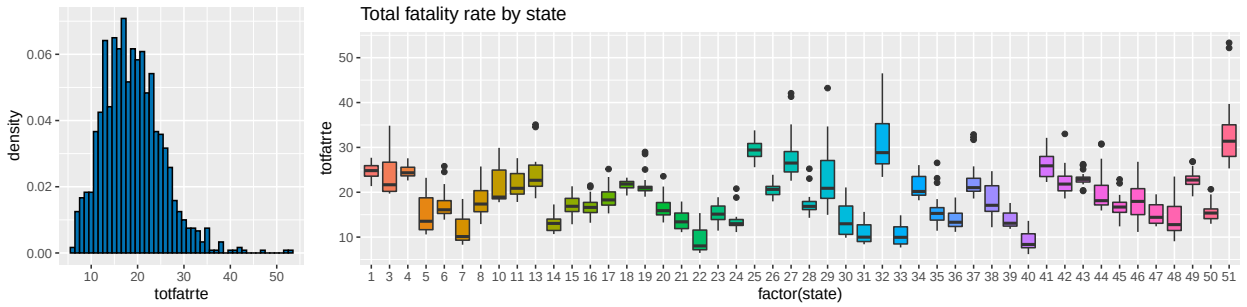


From the above graph `Fatality Rates by Year` in conjunction with `Total Miles Driver Per-Capita`, we can see that, without any assertion of causality, that overall fatality rates, as well as night and weekend fatality rates have steadily decreased over our time period of interest, while

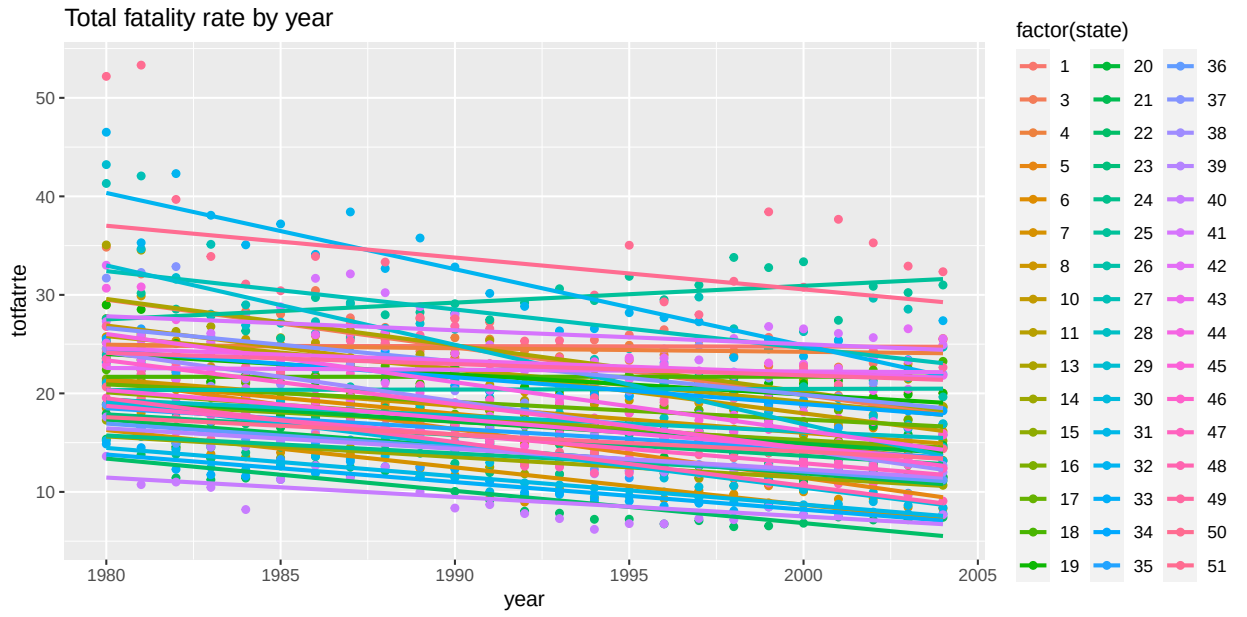
the number of miles driven per-capita has steadily increased. This provides substantial credibility to the hypothesis that driven has become safter, and should give us confidence in investigating potential causes.

Univariate analysis of the response variable

```
p1 <- ggplot(driving, aes(x = totfatrte)) +
      geom_histogram(aes(y = ..density..), binwidth = 1, fill="#0072B2", colour="black")
p2 <- ggplot(driving, aes(factor(state), totfatrte)) +
      geom_boxplot(aes(fill = factor(state))) + ggtitle('Total fatality rate by state')
ggarrange(p1, p2, ncol=2, nrow=1, widths=c(3, 9), legend="none")
```



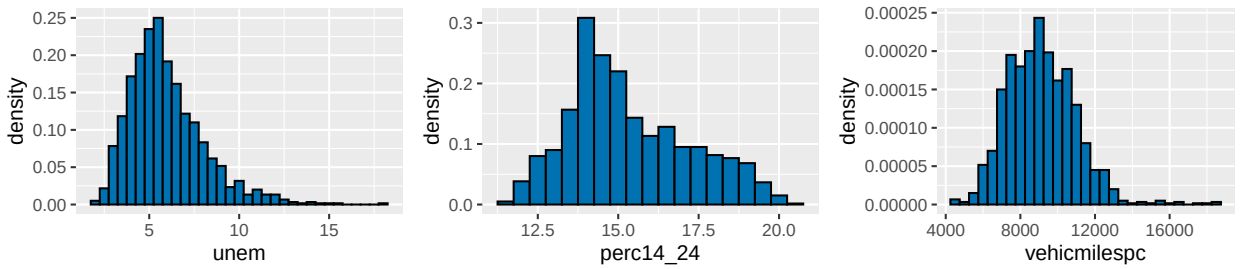

```
ggplot(driving, aes(x = year, y = totfatrte, color = factor(state))) +
  geom_point() + geom_smooth(method=lm, se=FALSE) + ggtitle('Total fatality rate by year')
```



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 box plot, we observed variated data variances across states. From the time plot grouped by state, we could see that for most states, the fatality rate trended to decrease from 1980 to 2004.

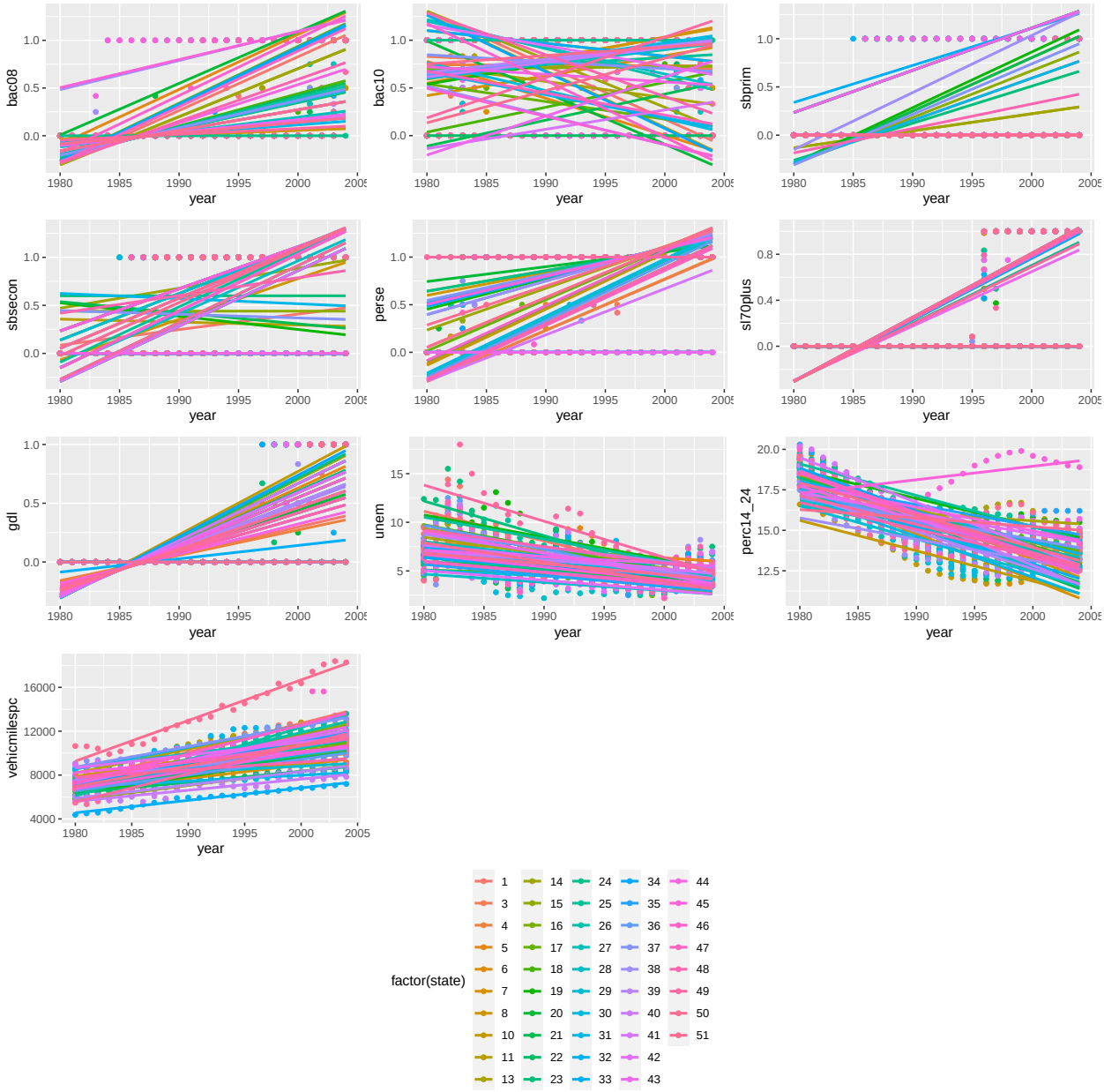
Univariate analysis of the explanatory variables

```
hist.unem <- ggplot(driving, aes(x = unem)) +
             geom_histogram(aes(y = ..density..), binwidth = 0.5, fill="#0072B2", colour="blac
hist.perc <- ggplot(driving, aes(x = perc14_24)) +
             geom_histogram(aes(y = ..density..), binwidth = 0.5, fill="#0072B2", colour="blac
hist.vmpc <- ggplot(driving, aes(x = vehicmilespc)) +
             geom_histogram(aes(y = ..density..), binwidth = 500, fill="#0072B2", colour="blac
ggarrange(hist.unem, hist.perc, hist.vmpc, ncol=3, nrow=1)
```



The distributions of unemployment rate, percent population aged 14 to 24 and vehicle miles driven per capita are all right-skewed. The skewness of *perc14_24* is more severe.

```
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=3, nrow=4, common.legend = TRUE, legend="botto
```



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 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.

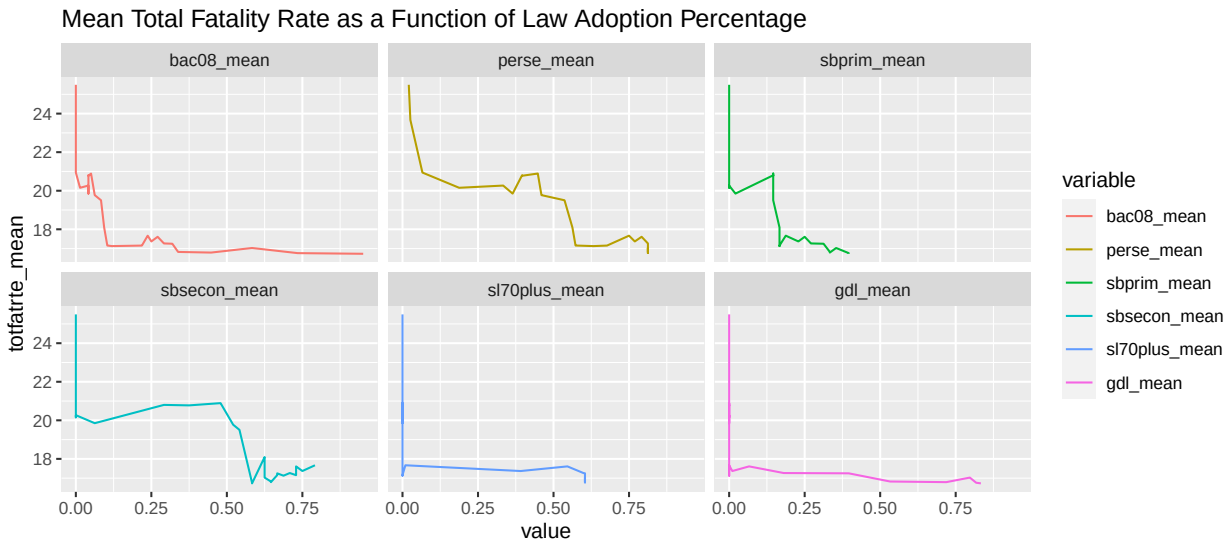
In most states, the enforcement of the "Per se" law trended to increase in the period. There are also some states where the law remained in effect or never in effect in the period. In a few states, the enforcement of speed limit of 70 and up trended to increase by time. Some states had never

enacted such high speed limit in the period. In fact, over 75% observations valued 0 in *sl70plus*. The enforcement of the graduated drivers license law trended to increase in a few states. 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 decreas while the trend vehicle miles traveled per capita was increase.

Initial analysis of the effcts of the law indicators on the fatality rate

```
# Take a look at overall correlations with specific laws. Since law variables are binary, we c
driving.law.means <- driving.restricted %>% group_by(year) %>%
  mutate(perc14_24 = perc14_24 / 100) %>% mutate(unem = unem / 100) %>%
  summarise_at(vars(totfatrte, bac08, perse, sbprim, sbsecon, sl70plus, gdl), list(mean=mean))
driving.adoption <- as.data.frame(driving.law.means)
driving.law.melt <- melt(driving.adoption, id.vars=c('year', 'totfatrte_mean'))
ggplot(driving.law.melt) +
  aes(value, totfatrte_mean, col=variable) + geom_line() + facet_wrap(~variable) +
  labs(title="Mean Total Fatality Rate as a Function of Law Adoption Percentage")
```

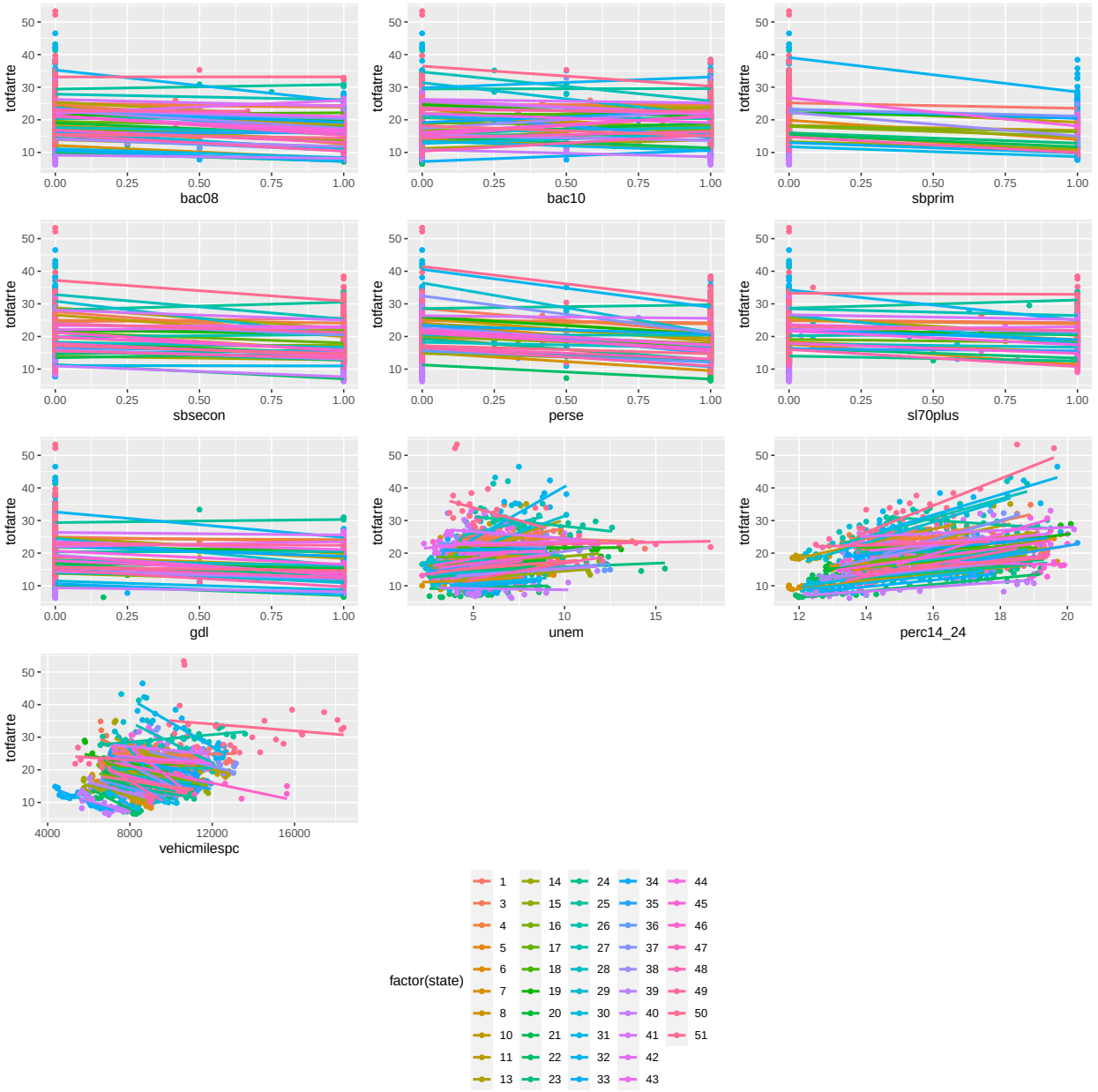


The above plot uses the mean of the binary indicator for a given law, provided by each state, as a proxy for the total adoption of the law across all states. For example, if no state has implemented the law in a given year, then the mean will be 0, if all states have adopted the law, it will be 1, otherwise it will be equal to $adoption\_rate(law) = \frac{1}{N} \cdot \sum_{i=1}^{N} I[law_i == 1]$, where I is the indicator function. If we plot total fatality rate as a function of law adoption rate, we might be able to observe something interesting.

In fact, we do observe some interesting correlation's between our total fatality rate and our proxy for adoption rate, indicating that there is reason to believe that adoption of certain laws, such as the `perse` law, does improve driver safety.

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=3, nrow=4, common.legend = TRUE, legend="bottom")
```

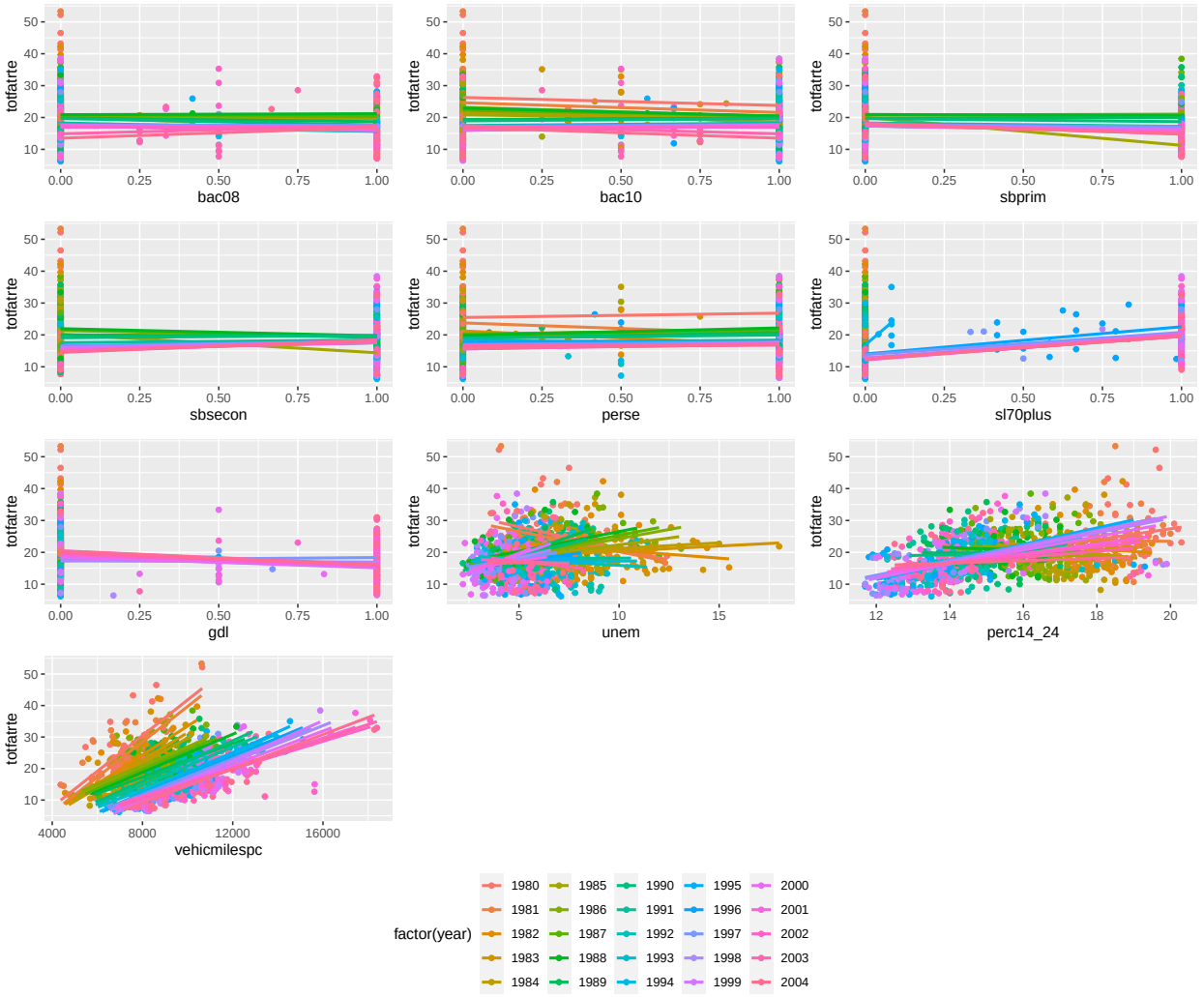


Within states, some negative correlation was observed between *bac08* and *totfatrte*. This suggests that for a given state, the enforcement of BAC limit of 0.08% would probably decrease the fatality rate. On the other hand, the correlation between *bac10* and *totfatrte* is not very clear. Negative correlations were observed between both the primary and the secondary seatbelt law and the fatality rate. The enforcement of the "Per se" laws was negatively correlated with the fatality rate for most states, with few exceptions. Negative effects were also observed on high speed limit (70 and up) and the enforcement of graduated drivers license law.

Virtually, for most states, the unemployment rate is positively correlated with the fatality rate with variated slopes among states. Similar effect was observed on the percent population aged 14 to 24. For vehicle miles driven per capita, the within states correlation seems to be negative for most states.

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(y
  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=3, nrow=4, common.legend = TRUE, legend="bottom")
```



Within a year, the correlation between *bac08* and *totfatrte* is not very obvious. On the other hand, some negative correlation was observed between *bac10* and *totfatrte*. The within years correlation between *sbprim* and *totfatrte* is also negative but that between *sbsecon* and *totfatrte* is mixed. Interestingly, some weakly positive effect was observed from *perse* while the positive effect was more obvious from *sl70plus*. This indicates that the effects of the enforcement of the "per se" law and high speed limit are complicated. Negative effect was observed from *gdl*.

In most years, *unem* were found positively correlated with *totfatrte* with variated regression slopes among years. Similar effects were also observed from *perc14_24* and *vehicmilespc*. For *vehicmilespc*, the regression slopes get decreased by year. It suggests that the positive effect of *vehicmilespc* on *totfatrte* shrinks over time. This may explain the seemingly 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? Did driving become safer over this period?

Total fatality rate is defined as the number of fatalities per 100,000 population. The rounded year

based means are as below.

```
# Average totfatrate by year
t(round(driving.means[c('year', 'totfatrte_mean')], 0))
```

```
##                [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]
## year           1980 1981 1982 1983 1984 1985 1986 1987 1988  1989  1990
## totfatrte_mean   25   24   21   20   20   20   21   21   21    20    20
##                [,12] [,13] [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21]
## year            1991  1992  1993  1994  1995  1996  1997  1998  1999  2000
## totfatrte_mean    18    17    17    17    18    17    18    17    17    17
##                [,22] [,23] [,24] [,25]
## year            2001  2002  2003  2004
## totfatrte_mean    17    17    17    17
```

```
# Extract dummies
driving.dummies <- driving %>% filter(year > 1980 ) %>%
  select(!matches('d80')) %>% select(matches("totfatrte") | matches("d[0-9]{2}"))
lm.yr.dummie.fit <- lm(driving.dummies)
summary(lm.yr.dummie.fit)
```

```
##
## Call:
## lm(formula = driving.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
```

The linear regression on the year dummies were estimated as above. 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?

The binary variables *bac08, bac10, perse, subprim, sbsecon, sl70plus, gdl* have value ranges from 0 to 1. No transformation is needed for these variables. The variables *perc14_24*, *unem* and *vehicmilespc* are continuous and normalized, but given that the skewness of *perc14_24* is more severed we decided to take logarithmic transformation on it.

```
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 + log(perc14_24) + unem +
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 + log(perc14_24) +
##     unem + vehicmilespc, data = driving)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.9146  -2.7322  -0.2732   2.2793  21.4225
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -6.258e+00  5.479e+00  -1.142 0.253641
## d81            -2.185e+00  8.273e-01  -2.641 0.008376 **
## d82            -6.615e+00  8.519e-01  -7.765 1.78e-14 ***
## d83            -7.425e+00  8.655e-01  -8.579  < 2e-16 ***
## d84            -5.887e+00  8.699e-01  -6.767 2.07e-11 ***
## d85            -6.526e+00  8.856e-01  -7.369 3.26e-13 ***
## d86            -5.900e+00  9.191e-01  -6.419 1.99e-10 ***
## d87            -6.418e+00  9.525e-01  -6.738 2.52e-11 ***
## d88            -6.645e+00  9.969e-01  -6.666 4.06e-11 ***
## d89            -8.124e+00  1.034e+00  -7.854 9.08e-15 ***
## d90            -9.011e+00  1.058e+00  -8.513  < 2e-16 ***
## d91            -1.112e+01  1.083e+00 -10.264  < 2e-16 ***
## d92            -1.293e+01  1.106e+00 -11.692  < 2e-16 ***
## d93            -1.278e+01  1.120e+00 -11.410  < 2e-16 ***
## d94            -1.241e+01  1.141e+00 -10.873  < 2e-16 ***
## d95            -1.200e+01  1.169e+00 -10.264  < 2e-16 ***
## d96            -1.392e+01  1.210e+00 -11.500  < 2e-16 ***
## d97            -1.430e+01  1.237e+00 -11.557  < 2e-16 ***
## d98            -1.508e+01  1.253e+00 -12.038  < 2e-16 ***
## d99            -1.513e+01  1.271e+00 -11.901  < 2e-16 ***
## d00            -1.549e+01  1.291e+00 -11.991  < 2e-16 ***
## d01            -1.623e+01  1.320e+00 -12.292  < 2e-16 ***
## d02            -1.677e+01  1.334e+00 -12.570  < 2e-16 ***
## d03            -1.707e+01  1.345e+00 -12.689  < 2e-16 ***
## d04            -1.676e+01  1.372e+00 -12.214  < 2e-16 ***
## bac08          -2.499e+00  5.375e-01  -4.649 3.72e-06 ***
## bac10          -1.423e+00  3.962e-01  -3.592 0.000342 ***
## perse          -6.189e-01  2.982e-01  -2.075 0.038194 *
## sbprim         -7.731e-02  4.908e-01  -0.158 0.874867
## sbsecon         6.741e-02  4.293e-01   0.157 0.875256
## sl70plus        3.344e+00  4.468e-01   7.485 1.41e-13 ***
## gdl            -4.258e-01  5.269e-01  -0.808 0.419230
## log(perc14_24)  2.125e+00  1.869e+00   1.137 0.255868
## unem            7.563e-01  7.788e-02   9.710  < 2e-16 ***
```

```
## vehicmilespc    2.923e-03  9.546e-05  30.618  < 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: 53.09 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?

```
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
            d94 + d95 + d96 + d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04 + bac08 + bac10 +
            perse + sbprim + sbsecon + sl70plus + gdl + log(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 + log(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.
## -8.4061129 -1.0255861 -0.0061669  0.9521453 14.8127642
##
## Coefficients:
##                    Estimate  Std. Error  t-value  Pr(>|t|)
## d81            -1.51819667  0.41262215  -3.6794 0.0002449 ***
## d82            -3.04113278  0.44050563  -6.9037 8.475e-12 ***
## d83            -3.52206046  0.45229479  -7.7871 1.557e-14 ***
## d84            -4.27513305  0.45749138  -9.3447 < 2.2e-16 ***
## d85            -4.74106246  0.47516016  -9.9778 < 2.2e-16 ***
## d86            -3.67173860  0.50519672  -7.2679 6.839e-13 ***
## d87            -4.30762611  0.54042417  -7.9708 3.866e-15 ***
## d88            -4.75795051  0.58508134  -8.1321 1.112e-15 ***
## d89            -6.10900872  0.62287680  -9.8077 < 2.2e-16 ***
## d90            -6.19960678  0.64768093  -9.5720 < 2.2e-16 ***
## d91            -6.88058635  0.66513726 -10.3446 < 2.2e-16 ***
## d92            -7.73040677  0.68720716 -11.2490 < 2.2e-16 ***
## d93            -8.04418827  0.70116232 -11.4726 < 2.2e-16 ***
## d94            -8.44753068  0.72051220 -11.7243 < 2.2e-16 ***
## d95            -8.19238737  0.74403807 -11.0107 < 2.2e-16 ***
## d96            -8.52552870  0.78558239 -10.8525 < 2.2e-16 ***
## d97            -8.62078789  0.81037726 -10.6380 < 2.2e-16 ***
## d98            -9.26142578  0.82494527 -11.2267 < 2.2e-16 ***
## d99            -9.38936854  0.83547075 -11.2384 < 2.2e-16 ***
## d00            -9.90897438  0.84707548 -11.6979 < 2.2e-16 ***
## d01            -9.55066012  0.86372249 -11.0576 < 2.2e-16 ***
## d02            -8.82873248  0.87328413 -10.1098 < 2.2e-16 ***
## d03            -8.86055501  0.88104728 -10.0568 < 2.2e-16 ***
## d04            -9.26313757  0.90236573 -10.2654 < 2.2e-16 ***
## bac08          -1.43849348  0.39402485  -3.6508 0.0002735 ***
## bac10          -1.06636724  0.26870967  -3.9685 7.695e-05 ***
## perse          -1.14750415  0.23381781  -4.9077 1.058e-06 ***
## sbprim         -1.23005317  0.34255376  -3.5908 0.0003439 ***
## sbsecon        -0.35058836  0.25206852  -1.3908 0.1645490
## sl70plus       -0.09096346  0.27048251  -0.3363 0.7367072
## gdl            -0.40741053  0.29235746  -1.3935 0.1637348
## log(perc14_24)  3.14688235  1.43687739   2.1901 0.0287242 *
## unem           -0.57035963  0.06056392  -9.4175 < 2.2e-16 ***
## vehicmilespc    0.00092993  0.00011141   8.3466 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    12134
## Residual Sum of Squares: 4531.5
## R-Squared:      0.62655
## Adj. R-Squared: 0.59949
## F-statistic: 55.1668 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  -2.4987093     -1.438493
## bac10  -1.4230058     -1.066367
## perse  -0.6188569     -1.147504
## sbprim -0.0773098     -1.230053
```

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*?

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

```
rnd_e <- plm(data = driving.panel,
          totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89 + d90 + d91 + d92 + d
            d94 + d95 + d96 + d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04 + bac08 + bac10 +
            perse + sbprim + sbsecon + sl70plus + gdl + log(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 + log(perc14_24) +
##     unem + vehicmilespc, data = driving.panel, model = "random")
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Effects:
##                  var std.dev share
## idiosyncratic 4.053   2.013 0.328
## individual    8.289   2.879 0.672
## theta: 0.8615
##
## Residuals:
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -8.23654 -1.14421 -0.15409  0.92221 16.46238
##
## Coefficients:
##                    Estimate  Std. Error  z-value  Pr(>|z|)
## (Intercept)     1.1387e+01  4.3834e+00   2.5978 0.0093824 **
## d81            -1.5574e+00  4.2772e-01  -3.6412 0.0002714 ***
## d82            -3.2618e+00  4.5580e-01  -7.1564 8.285e-13 ***
## d83            -3.7679e+00  4.6780e-01  -8.0546 7.975e-16 ***
## d84            -4.3955e+00  4.7310e-01  -9.2909 < 2.2e-16 ***
## d85            -4.8836e+00  4.9091e-01  -9.9482 < 2.2e-16 ***
## d86            -3.8499e+00  5.2146e-01  -7.3829 1.548e-13 ***
## d87            -4.5150e+00  5.5695e-01  -8.1067 5.203e-16 ***
## d88            -4.9864e+00  6.0204e-01  -8.2825 < 2.2e-16 ***
## d89            -6.3655e+00  6.4024e-01  -9.9423 < 2.2e-16 ***
## d90            -6.5216e+00  6.6517e-01  -9.8044 < 2.2e-16 ***
```

```
## d91            -7.2826e+00  6.8305e-01 -10.6619 < 2.2e-16 ***
## d92            -8.2120e+00  7.0514e-01 -11.6459 < 2.2e-16 ***
## d93            -8.5090e+00  7.1925e-01 -11.8303 < 2.2e-16 ***
## d94            -8.8791e+00  7.3892e-01 -12.0163 < 2.2e-16 ***
## d95            -8.6316e+00  7.6276e-01 -11.3163 < 2.2e-16 ***
## d96            -9.0347e+00  8.0494e-01 -11.2241 < 2.2e-16 ***
## d97            -9.1526e+00  8.3003e-01 -11.0267 < 2.2e-16 ***
## d98            -9.8236e+00  8.4456e-01 -11.6316 < 2.2e-16 ***
## d99            -9.9650e+00  8.5528e-01 -11.6512 < 2.2e-16 ***
## d00            -1.0485e+01  8.6720e-01 -12.0908 < 2.2e-16 ***
## d01            -1.0212e+01  8.8403e-01 -11.5515 < 2.2e-16 ***
## d02            -9.5781e+00  8.9350e-01 -10.7197 < 2.2e-16 ***
## d03            -9.6253e+00  9.0148e-01 -10.6772 < 2.2e-16 ***
## d04            -9.9966e+00  9.2331e-01 -10.8269 < 2.2e-16 ***
## bac08          -1.5709e+00  4.0367e-01  -3.8916 9.958e-05 ***
## bac10          -1.1421e+00  2.7593e-01  -4.1392 3.486e-05 ***
## perse          -1.0900e+00  2.3871e-01  -4.5659 4.973e-06 ***
## sbprim         -1.1792e+00  3.5130e-01  -3.3566 0.0007890 ***
## sbsecon        -3.4854e-01  2.6016e-01  -1.3397 0.1803283
## sl70plus        4.1804e-03  2.7895e-01   0.0150 0.9880432
## gdl            -3.8188e-01  3.0229e-01  -1.2633 0.2064959
## log(perc14_24)  3.2485e+00  1.4695e+00   2.2106 0.0270629 *
## unem           -4.9115e-01  6.1827e-02  -7.9439 1.959e-15 ***
## vehicmilespc    1.1648e-03  1.1022e-04  10.5682 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    12834
## Residual Sum of Squares: 5075.1
## R-Squared:      0.60455
## Adj. R-Squared: 0.59301
## Chisq: 1781.05 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:

```
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      -1.438493      -1.570937
## bac10      -1.066367      -1.142110
## perse      -1.147504      -1.089952
## sbprim     -1.230053      -1.179175
```

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.

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

```
##
##  Hausman Test
##
## data:  totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89 +  ...
## chisq = 150.92, df = 34, p-value < 2.2e-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.

```
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.