Statistical Methods for Discrete Response, Time Series, and Panel Data (W271): Lab 4

Heather Feinstein, Ryan Delgado, April Kim

Fall 2018

Instructions:

- **Due Date: 12/11/2018 (11:59 p.m. Pacific Time)**
- **Page limit of the pdf report: 20 (not include title and the table of content page**
- Use the margin, linespace, and font size specification below:
  - fontsize=11pt
  - margin=1in
  - line_spacing=single
- Submission:
  - Each group makes one submission to Github; please have one of your team members made the submission
  - Submit 2 files:
    1. A pdf file including the details of your analysis and all the R codes used to produce the analysis. Please do not suppress the codes in your pdf file.
    2. R markdown file used to produce the pdf file
  - Use the following file-naming convensation; fail to do so will receive 10% reduction in the grade:
    - FirstNameLastName1_FirstNameLastName2_FirstNameLastName3_LabNumber.fileExtension
    - For example, if you have three students in the group for Lab Z, and their names are Gerard Kelley, Steve Yang, and Jeffrey Yau, then you should name your file the following
      - GerardKelley_SteveYang_JeffreyYau_LabZ.Rmd
      - GerardKelley_SteveYang_JeffreyYau_LabZ.pdf
  - Although it sounds obvious, please write the name of each members of your group on page 1 of your pdf and Rmd files.
- This lab can be completed in a group of up to 3 students in your session. Students are encouraged to work in a group for the lab.
- For statistical methods that we cover in this course, use only the R libraries and functions that are covered in this course. If you use libraries and functions for statistical modeling that we have not covered, you have to provide (1) explanation of why such libraries and functions are used instead and (2) reference to the suppressWarnings(suppressMessages(library documentation. Lacking the explanation and reference to the documentation will result in a score of zero for the corresponding question.
- Students are expected to act with regards to UC Berkeley Academic Integrity.

Description of the Lab

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

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

Exercises:

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

```
# Insert the function to *tidy up* the code when they are printed out
suppressWarnings(suppressMessages(library(knitr)))
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)

# Load libraries
suppressWarnings(suppressMessages(library(car)))
suppressWarnings(suppressMessages(library(dplyr)))
suppressWarnings(suppressMessages(library(Hmisc)))
suppressWarnings(suppressMessages(library(ggplot2)))
suppressWarnings(suppressMessages(library(lattice)))
suppressWarnings(suppressMessages(library(plm)))
suppressWarnings(suppressMessages(library(plyr)))
```

```
load("driving.RData")

head(data, 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
## 2 1981     1    1    0    0    0      0        0     18       0   0     1
## 3 1982     1    1    0    0    0      0        0     18       0   0     1
## 4 1983     1    1    0    0    0      0        0     18       0   0     1
## 5 1984     1    1    0    0    0      0        0     18       0   0     1
##   bac08 perse totfat nghtfat wkndfat totfatpvm nghtfatpvm wkndfatpvm
## 1     0     0    940     422     236      3.20      1.437      0.803
## 2     0     0    933     434     248      3.35      1.558      0.890
## 3     0     0    839     376     224      2.81      1.259      0.750
## 4     0     0    930     397     223      3.00      1.281      0.719
## 5     0     0    932     421     237      2.83      1.278      0.720
##   statepop totfatrte nghtfatrte wkndfatrte vehicmiles unem perc14_24
## 1  3893888     24.14      10.84       6.06   29.37500  8.8      18.9
## 2  3918520     24.07      11.08       6.33   27.85200 10.7      18.7
## 3  3925218     21.37       9.58       5.71   29.85765 14.4      18.4
```

```
## 4  3934109     23.64      10.09       5.67   31.00000 13.7      18.0
## 5  3951834     23.58      10.65       6.00   32.93286 11.1      17.6
##   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
##   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
```

```
desc
```

```
##        variable                                        label
## 1          year                             1980 through 2004
## 2         state            48 continental states, alphabetical
## 3          sl55                               speed limit == 55
## 4          sl65                               speed limit == 65
## 5          sl70                               speed limit == 70
## 6          sl75                               speed limit == 75
## 7        slnone                                  no speed limit
## 8      seatbelt     =0 if none, =1 if primary, =2 if secondary
## 9        minage                              minimum drinking age
## 10      zerotol                                zero tolerance law
## 11          gdl                    graduated drivers license law
## 12        bac10                          blood alcohol limit .10
## 13        bac08                          blood alcohol limit .08
## 14        perse administrative license revocation (per se law)
## 15       totfat                          total traffic fatalities
## 16      nghtfat                       total nighttime fatalities
## 17      wkndfat                          total weekend fatalities
## 18    totfatpvm         total fatalities per 100 million miles
## 19   nghtfatpvm     nighttime fatalities per 100 million miles
## 20   wkndfatpvm       weekend fatalities per 100 million miles
## 21     statepop                                state population
## 22    totfatrte        total fatalities per 100,000 population
## 23   nghtfatrte    nighttime fatalities per 100,000 population
## 24   wkndfatrte       weekend accidents per 100,000 population
## 25   vehicmiles               vehicle miles traveled, billions
## 26         unem                     unemployment rate, percent
## 27    perc14_24          percent population aged 14 through 24
## 28     sl70plus                              sl70 + sl75 + slnone
## 29       sbprim                      =1 if primary seatbelt law
## 30      sbsecon                    =1 if secondary seatbelt law
## 31          d80                                =1 if year == 1980
## 32          d81
## 33          d82
## 34          d83
## 35          d84
## 36          d85
```

```
## 37          d86
## 38          d87
## 39          d88
## 40          d89
## 41          d90
## 42          d91
## 43          d92
## 44          d93
## 45          d94
## 46          d95
## 47          d96
## 48          d97
## 49          d98
## 50          d99
## 51          d00
## 52          d01
## 53          d02
## 54          d03
## 55          d04                              =1 if year == 2004
## 56 vehicmilespc
```

```
# each state has exactly 25 obsearvations -- panel is
# balanced
min(table(data$state))
```

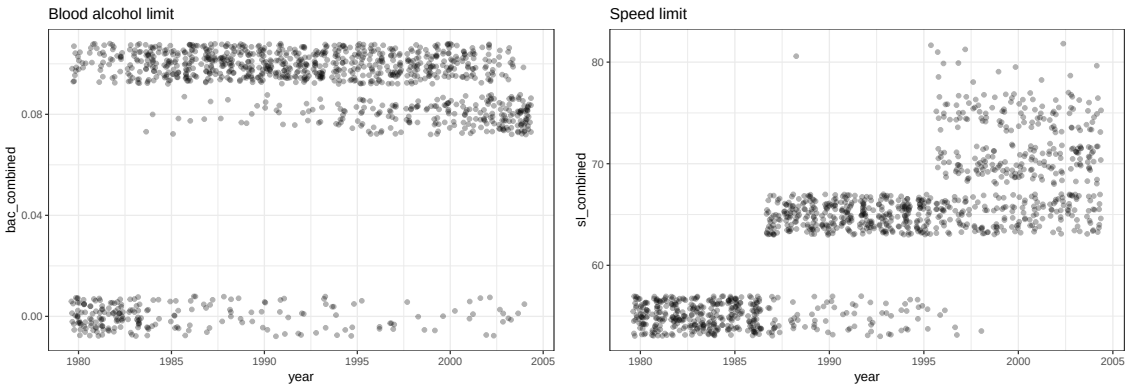
```
## [1] 25
```

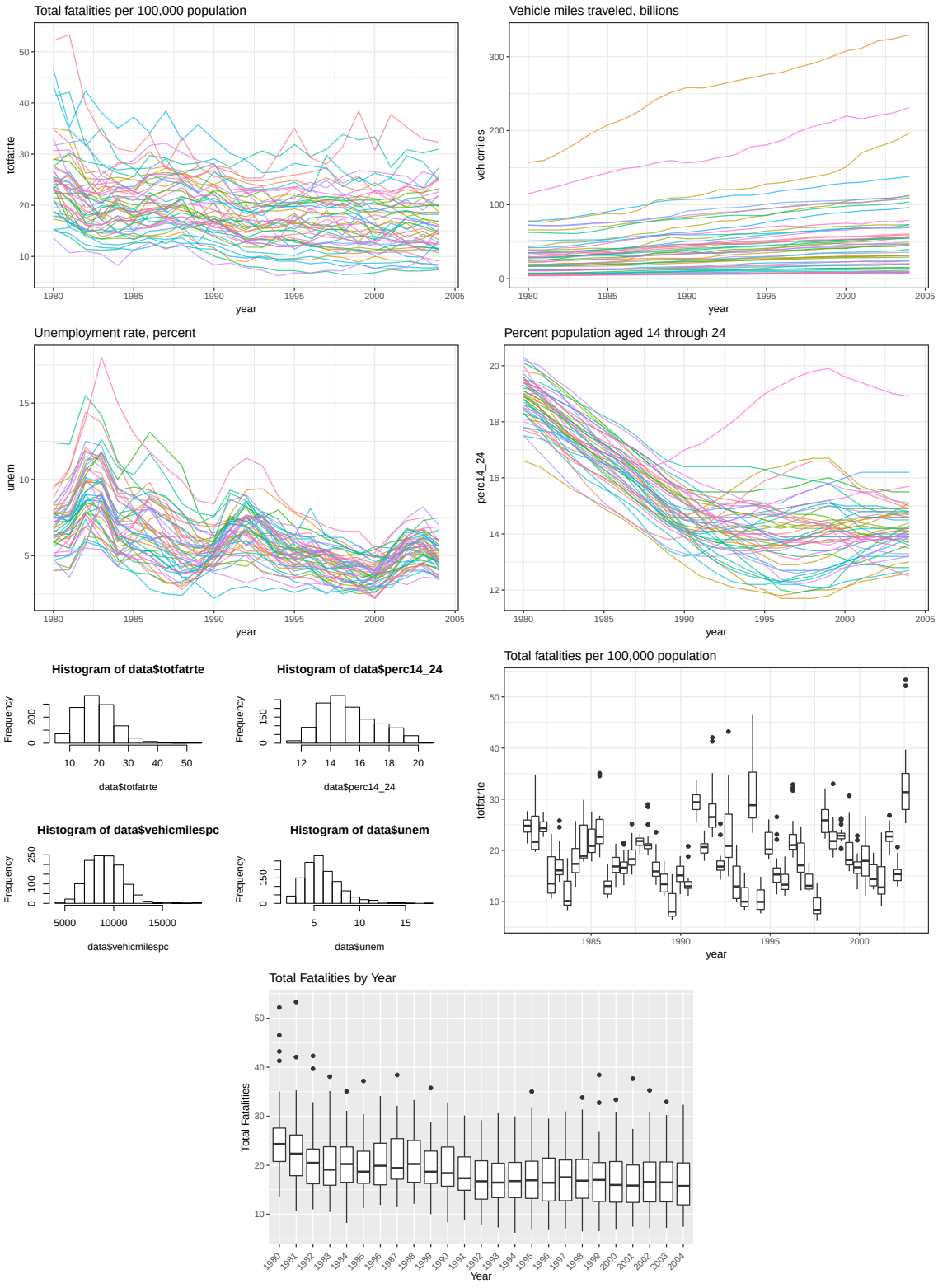
```
max(table(data$state))
```

```
## [1] 25
```

```
## viz variables to use every state throughout the years:
## totfartre, perc14_24, vehicmilespc and unem

# distribution across diff state: seatbelt, zerotol, gdl,
# perse, bac08, bac01 and slXX variables

# if '1' in bac10, set as 0.1; else set as '0' or '1' in
# bac08
data$bac_combined <- ifelse(round(data$bac10) > 0, 0.1, 0.08 *
    round(data$bac08))
data$sl_combined <- ifelse(round(data$sl55) > 0, 55, ifelse(round(data$sl65) >
    0, 65, ifelse(round(data$sl70) > 0, 70, ifelse(round(data$sl75) >
    0, 75, 80))))

ggplot(data, aes(x = year, y = bac_combined, group = state)) +
    geom_jitter(alpha = 0.3) + theme_bw() + ggtitle("Blood alcohol limit")
ggplot(data, aes(x = year, y = sl_combined, group = state)) +
    geom_jitter(alpha = 0.3) + theme_bw() + ggtitle("Speed limit")

# totfatrte: total fatalities per 100,000 population
ggplot(data, aes(x = year, y = totfatrte, colour = as.factor(state))) +
    geom_line(alpha = 0.7, show.legend = F) + ggtitle("Total fatalities per 100,000 population") +
```

```
    theme_bw()

# vehicmiles vehicle miles traveled, billions
ggplot(data, aes(x = year, y = vehicmiles, colour = as.factor(state))) +
    geom_line(alpha = 0.7, show.legend = F) + ggtitle("Vehicle miles traveled, billions") +
    theme_bw()

# unem unemployment rate, percent
ggplot(data, aes(x = year, y = unem, colour = as.factor(state))) +
    geom_line(alpha = 0.7, show.legend = F) + ggtitle("Unemployment rate, percent") +
    theme_bw()

# perc14_24 percent population aged 14 through 24
ggplot(data, aes(x = year, y = perc14_24, colour = as.factor(state))) +
    geom_line(alpha = 0.7, show.legend = F) + ggtitle("Percent population aged 14 through 24") +
    theme_bw()

par(mfrow = c(2, 2))
hist(data$totfatrte)
hist(data$perc14_24)
hist(data$vehicmilespc)
hist(data$unem)

ggplot(data, aes(x = year, y = totfatrte, group = as.factor(state))) +
    geom_boxplot() + ggtitle("Total fatalities per 100,000 population") +
    theme_bw()

# ryan's plot
ggplot(data, aes(as.factor(year), totfatrte)) + geom_boxplot() +
    ggtitle("Total Fatalities by Year") + xlab("Year") + ylab("Total Fatalities") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1.2))
```





- blood alcohol limit was 0.1 in the 80s; starting in mid 80s reduced to 0.08 as well as set zero to 0.08
- speed limit increased with time
- decrease in total fatality from 1980 to 1985; remained roughly constant post 1985

- positive linear trend in vehicles miles traveled
- increase in unemployment rate in early 80s followed by decrease until early 90s; steady decrease from mid 90s
- rapid decrease in percent population aged 14 through 24 until early 90s; remained constant afterwards
- observe some degree of skewness and may need to transform (histogram)
- large differences in distribution of total fatality rate between states

2. 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.

totfatrte is defined as "nighttime fatalities per 100,000 population"

```
ddply(data, .(year), summarize, Total = mean(totfatrte))
```

```
##    year    Total
## 1  1980 25.49458
## 2  1981 23.67021
## 3  1982 20.94250
## 4  1983 20.15292
## 5  1984 20.26750
## 6  1985 19.85146
## 7  1986 20.80042
## 8  1987 20.77479
## 9  1988 20.89167
## 10 1989 19.77229
## 11 1990 19.50521
## 12 1991 18.09479
## 13 1992 17.15792
## 14 1993 17.12771
## 15 1994 17.15521
## 16 1995 17.66854
## 17 1996 17.36938
## 18 1997 17.61062
## 19 1998 17.26542
## 20 1999 17.25042
## 21 2000 16.82562
## 22 2001 16.79271
## 23 2002 17.02958
## 24 2003 16.76354
## 25 2004 16.72896
```

```
mod1 <- lm(totfatrte ~ factor(year), data = data)
summary(mod1)
```

```
##
## Call:
```

```
## lm(formula = totfatrte ~ factor(year), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9302  -4.3468  -0.7305   3.7488  29.6498
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       25.4946     0.8671  29.401  < 2e-16 ***
## factor(year)1981  -1.8244     1.2263  -1.488 0.137094
## factor(year)1982  -4.5521     1.2263  -3.712 0.000215 ***
## factor(year)1983  -5.3417     1.2263  -4.356 1.44e-05 ***
## factor(year)1984  -5.2271     1.2263  -4.263 2.18e-05 ***
## factor(year)1985  -5.6431     1.2263  -4.602 4.64e-06 ***
## factor(year)1986  -4.6942     1.2263  -3.828 0.000136 ***
## factor(year)1987  -4.7198     1.2263  -3.849 0.000125 ***
## factor(year)1988  -4.6029     1.2263  -3.754 0.000183 ***
## factor(year)1989  -5.7223     1.2263  -4.666 3.42e-06 ***
## factor(year)1990  -5.9894     1.2263  -4.884 1.18e-06 ***
## factor(year)1991  -7.3998     1.2263  -6.034 2.14e-09 ***
## factor(year)1992  -8.3367     1.2263  -6.798 1.68e-11 ***
## factor(year)1993  -8.3669     1.2263  -6.823 1.43e-11 ***
## factor(year)1994  -8.3394     1.2263  -6.800 1.66e-11 ***
## factor(year)1995  -7.8260     1.2263  -6.382 2.51e-10 ***
## factor(year)1996  -8.1252     1.2263  -6.626 5.25e-11 ***
## factor(year)1997  -7.8840     1.2263  -6.429 1.86e-10 ***
## factor(year)1998  -8.2292     1.2263  -6.711 3.01e-11 ***
## factor(year)1999  -8.2442     1.2263  -6.723 2.77e-11 ***
## factor(year)2000  -8.6690     1.2263  -7.069 2.67e-12 ***
## factor(year)2001  -8.7019     1.2263  -7.096 2.21e-12 ***
## factor(year)2002  -8.4650     1.2263  -6.903 8.32e-12 ***
## factor(year)2003  -8.7310     1.2263  -7.120 1.88e-12 ***
## factor(year)2004  -8.7656     1.2263  -7.148 1.54e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.008 on 1175 degrees of freedom
## Multiple R-squared:  0.1276, Adjusted R-squared:  0.1098
## F-statistic: 7.164 on 24 and 1175 DF,  p-value: < 2.2e-16
```

This model suggests that driving got safer over the time period as each year has an increasing beta and all years are significant after 1981.

3. 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.)

```r
mod2 <- lm(totfatrte ~ factor(year) + bac08 + bac10 + perse +
    sbprim + sbsecon + sl70plus + gdl + perc14_24 + gdl + perc14_24 +
    unem + vehicmilespc, data = data)

summary(mod2)
```

```
##
## Call:
## lm(formula = totfatrte ~ factor(year) + bac08 + bac10 + perse +
##     sbprim + sbsecon + sl70plus + gdl + perc14_24 + gdl + perc14_24 +
##     unem + vehicmilespc, data = data)
##
## 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
## factor(year)1981 -2.175e+00  8.276e-01  -2.629 0.008686 **
## factor(year)1982 -6.596e+00  8.534e-01  -7.729 2.33e-14 ***
## factor(year)1983 -7.397e+00  8.690e-01  -8.512  < 2e-16 ***
## factor(year)1984 -5.850e+00  8.763e-01  -6.676 3.79e-11 ***
## factor(year)1985 -6.483e+00  8.948e-01  -7.245 7.82e-13 ***
## factor(year)1986 -5.853e+00  9.307e-01  -6.289 4.52e-10 ***
## factor(year)1987 -6.367e+00  9.670e-01  -6.585 6.87e-11 ***
## factor(year)1988 -6.592e+00  1.014e+00  -6.502 1.17e-10 ***
## factor(year)1989 -8.071e+00  1.053e+00  -7.667 3.68e-14 ***
## factor(year)1990 -8.959e+00  1.077e+00  -8.319 2.46e-16 ***
## factor(year)1991 -1.107e+01  1.101e+00 -10.052  < 2e-16 ***
## factor(year)1992 -1.288e+01  1.123e+00 -11.473  < 2e-16 ***
## factor(year)1993 -1.273e+01  1.136e+00 -11.204  < 2e-16 ***
## factor(year)1994 -1.236e+01  1.157e+00 -10.685  < 2e-16 ***
## factor(year)1995 -1.195e+01  1.184e+00 -10.098  < 2e-16 ***
## factor(year)1996 -1.388e+01  1.223e+00 -11.343  < 2e-16 ***
## factor(year)1997 -1.426e+01  1.250e+00 -11.408  < 2e-16 ***
## factor(year)1998 -1.504e+01  1.265e+00 -11.886  < 2e-16 ***
## factor(year)1999 -1.509e+01  1.284e+00 -11.750  < 2e-16 ***
## factor(year)2000 -1.544e+01  1.305e+00 -11.831  < 2e-16 ***
## factor(year)2001 -1.618e+01  1.334e+00 -12.131  < 2e-16 ***
## factor(year)2002 -1.672e+01  1.348e+00 -12.406  < 2e-16 ***
## factor(year)2003 -1.702e+01  1.359e+00 -12.521  < 2e-16 ***
## factor(year)2004 -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
```

bac8 is blood alcohol limit .08 bac10 is blood alcohol limit .10

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

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

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

7. 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?