Lab 3: Panel Models

US Traffic Fatalities: 1980 - 2004

Contents


```
theme_set(theme_minimal())
```

1 U.S. traffic fatalities: 1980-2004

In this lab, we are asking you to answer the following **causal** question:

> **"Do changes in traffic laws affect traffic fatalities?"**

To answer this question, please complete the tasks specified below using the data provided in `data/driving.Rdata`. This data 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 also provided in the dataset.

```
setwd("/home/rstudio/kumarn/MIDS/w271/Lab-3")
load(file="./data/driving.RData")
```

```r
# quick view of the data and its parameters
glimpse(data)
```

```
## Rows: 1,200
## Columns: 56
## $ year         <int> 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 198~
## $ state        <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ sl55         <dbl> 1.000, 1.000, 1.000, 1.000, 1.000, 1.000, 1.000, 0.542, 0~
## $ sl65         <dbl> 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0.458, 1~
## $ sl70         <dbl> 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0~
## $ sl75         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ slnone       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ seatbelt     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2, 2, 2, 2, 2, 2, ~
## $ minage       <dbl> 18, 18, 18, 18, 18, 20, 21, 21, 21, 21, 21, 21, 21, 21, 2~
## $ zerotol      <dbl> 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0~
## $ gdl          <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.0~
## $ bac10        <dbl> 1.000, 1.000, 1.000, 1.000, 1.000, 1.000, 1.000, 1.000, 1~
## $ bac08        <dbl> 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0~
## $ perse        <dbl> 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0~
## $ totfat       <int> 940, 933, 839, 930, 932, 882, 1080, 1111, 1024, 1029, 112~
## $ nghtfat      <int> 422, 434, 376, 397, 421, 358, 500, 499, 423, 418, 466, 47~
## $ wkndfat      <int> 236, 248, 224, 223, 237, 224, 279, 300, 226, 247, 271, 27~
## $ totfatpvm    <dbl> 3.200, 3.350, 2.810, 3.000, 2.830, 2.510, 3.177, 2.970, 2~
## $ nghtfatpvm   <dbl> 1.437, 1.558, 1.259, 1.281, 1.278, 1.019, 1.471, 1.334, 1~
## $ wkndfatpvm   <dbl> 0.803, 0.890, 0.750, 0.719, 0.720, 0.637, 0.821, 0.802, 0~
## $ statepop     <int> 3893888, 3918520, 3925218, 3934109, 3951834, 3972527, 399~
## $ totfatrte    <dbl> 24.14, 24.07, 21.37, 23.64, 23.58, 22.20, 27.08, 27.67, 2~
## $ nghtfatrte   <dbl> 10.84, 11.08, 9.58, 10.09, 10.65, 9.01, 12.53, 12.43, 10.~
## $ wkndfatrte   <dbl> 6.060000, 6.330000, 5.710000, 5.670000, 6.000000, 5.64000~
## $ vehicmiles   <dbl> 29.37500, 27.85200, 29.85765, 31.00000, 32.93286, 35.1394~
## $ unem         <dbl> 8.8, 10.7, 14.4, 13.7, 11.1, 8.9, 9.8, 7.8, 7.2, 7.0, 6.9~
## $ perc14_24    <dbl> 18.9, 18.7, 18.4, 18.0, 17.6, 17.3, 17.0, 16.6, 16.2, 15.~
## $ sl70plus     <dbl> 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0~
## $ sbprim       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ sbsecon      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, ~
## $ d80          <int> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ d81          <int> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ d82          <int> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ d83          <int> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ d84          <int> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ d85          <int> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ d86          <int> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ d87          <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ d88          <int> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ d89          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ d90          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ d91          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, ~
## $ d92          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, ~
## $ d93          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, ~
## $ d94          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, ~
## $ d95          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ~
## $ d96          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, ~
## $ d97          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, ~
```

```
## $ d98          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, ~
## $ d99          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ d00          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ d01          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ d02          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ d03          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ d04          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ vehicmilespc <dbl> 7543.874, 7107.785, 7606.622, 7879.802, 8333.562, 8845.61~
```

```
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
```

```
# get the column names in character format
desc$label <- as.character(desc$label)

# save the data into a data frame
traffic_df <- data
```

2 (30 points, total) Build and Describe the Data

1. (5 points) Load the data and produce useful features. Specifically:
   - Produce a new variable, called `speed_limit` that re-encodes the data that is in `sl55`, `sl65`, `sl70`, `sl75`, and `slnone`;
   - Produce a new variable, called `year_of_observation` that re-encodes the data that is in `d80`, `d81`, ... , `d04`.
   - Produce a new variable for each of the other variables that are one-hot encoded (i.e. `bac*` variable series).
   - Rename these variables to sensible names that are legible to a reader of your analysis. For example, the dependent variable as provided is called, `totfatrte`. Pick something more sensible, like, `total_fatalities_rate`. There are few enough of these variables to change, that you should change them for all the variables in the data. (You will thank yourself later.)

```
## variable names
names(traffic_df)
```

```
##  [1] "year"         "state"        "sl55"         "sl65"         "sl70"
##  [6] "sl75"         "slnone"       "seatbelt"     "minage"       "zerotol"
## [11] "gdl"          "bac10"        "bac08"        "perse"        "totfat"
## [16] "nghtfat"      "wkndfat"      "totfatpvm"    "nghtfatpvm"   "wkndfatpvm"
## [21] "statepop"     "totfatrte"    "nghtfatrte"   "wkndfatrte"   "vehicmiles"
## [26] "unem"         "perc14_24"    "sl70plus"     "sbprim"       "sbsecon"
## [31] "d80"          "d81"          "d82"          "d83"          "d84"
## [36] "d85"          "d86"          "d87"          "d88"          "d89"
## [41] "d90"          "d91"          "d92"          "d93"          "d94"
## [46] "d95"          "d96"          "d97"          "d98"          "d99"
## [51] "d00"          "d01"          "d02"          "d03"          "d04"
## [56] "vehicmilespc"
```

```
# dimension of the data frame
dim(traffic_df)
```

```
## [1] 1200   56
```

```r
# check if the panel data is balanced
traff_table <- traffic_df %>% dplyr::select(year, state) %>% table()

# compute the row sums from the table above
val <- rowSums(traff_table)
val
```

```
## 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995
##   48   48   48   48   48   48   48   48   48   48   48   48   48   48   48   48
## 1996 1997 1998 1999 2000 2001 2002 2003 2004
##   48   48   48   48   48   48   48   48   48
```

```r
unbalanced_data <- 0

# in a loop check if any of the value is less than 48 (num. unique states)
for (i in val) {
  if (i != length(unique(traffic_df$state))) {
    state_code <- states_df[states_df$index == i, ]$state
    cat(paste("State", state_code, " at index", i, "is not balanced"))
    cat("------------------------\n")
    unbalanced_data <- unbalanced_data + 1
  }
}

if (unbalanced_data) {
  cat(pastes("-There are ", unbalanced_data, "states with unbalanced data-\n"))
} else {
  cat("-------The data set is balanced-------\n")
}
```

```
## -------The data set is balanced-------
```

```r
# check for gaps in the time series of each state
traffic_df %>% is.pconsecutive(index=c("state", "year"))
```

```
##    1    3    4    5    6    7    8   10   11   13   14   15   16   17   18   19
## TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##   20   21   22   23   24   25   26   27   28   29   30   31   32   33   34   35
## TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##   36   37   38   39   40   41   42   43   44   45   46   47   48   49   50   51
## TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

```r
if (sum(traffic_df %>% is.pconsecutive(index=c("state", "year")) - TRUE)) {
  cat("-----Gaps exist in data; check for gaps-----\n")
} else {
  cat("-----There are no gaps in the data------\n")
}
```

```
## -----There are no gaps in the data------
```

```r
# new variable, speed_limit; re-encode columns sl55, sl60, sl65, sl70, sl75, slnone
traffic_df$speed_limit <- ifelse(traffic_df$sl55 >= 0.5, 55,
                                 ifelse(traffic_df$sl60 >= 0.5, 60,
                                     ifelse(traffic_df$sl65 >= 0.5, 65,
                                         ifelse(traffic_df$sl70 >= 0.5, 70,
                                             ifelse(traffic_df$sl75 >= 0.5, 75, 0
                                                    )))))
```

```
# new variable, year_of_observation; re-encode columns d80, ..., d04
traffic_df$year_of_observation <- ifelse(traffic_df$d80 == 1, 1980,
                                  ifelse(traffic_df$d81 == 1, 1981,
                                  ifelse(traffic_df$d82 == 1, 1982,
                                  ifelse(traffic_df$d83 == 1, 1983,
                                  ifelse(traffic_df$d84 == 1, 1984,
                                  ifelse(traffic_df$d85 == 1, 1985,
                                  ifelse(traffic_df$d86 == 1, 1986,
                                  ifelse(traffic_df$d87 == 1, 1987,
                                  ifelse(traffic_df$d88 == 1, 1988,
                                  ifelse(traffic_df$d89 == 1, 1989,
                                  ifelse(traffic_df$d90 == 1, 1990,
                                  ifelse(traffic_df$d91 == 1, 1991,
                                  ifelse(traffic_df$d92 == 1, 1992,
                                  ifelse(traffic_df$d93 == 1, 1993,
                                  ifelse(traffic_df$d94 == 1, 1994,
                                  ifelse(traffic_df$d95 == 1, 1995,
                                  ifelse(traffic_df$d96 == 1, 1996,
                                  ifelse(traffic_df$d97 == 1, 1997,
                                  ifelse(traffic_df$d98 == 1, 1998,
                                  ifelse(traffic_df$d99 == 1, 1999,
                                  ifelse(traffic_df$d00 == 1, 2000,
                                  ifelse(traffic_df$d01 == 1, 2001,
                                  ifelse(traffic_df$d02 == 1, 2002,
                                  ifelse(traffic_df$d03 == 1, 2003,
                                         2004))))))))))))))))))))))))))

# new variable for each of the other one-hot encoded variables (bac* variables)
unique(traffic_df$bac10)
```

```
##  [1] 1.000 0.583 0.000 0.417 0.667 0.750 0.833 0.500 0.250 0.333
```

```
traffic_df$blood_alc_lim_10 <- ifelse(traffic_df$bac10 >= 0.5, 1, 0)
unique(traffic_df$bac08)
```

```
## [1] 0.000 0.417 1.000 0.333 0.500 0.250 0.750 0.667
```

```
traffic_df$blood_alc_lim_08 <- ifelse(traffic_df$bac08 >= 0.5, 1, 0)

# rename columns to reflect a meaningful description of the column
traffic_df <- traffic_df %>% rename( c(
  "gdl" = "grad_drivers_lic_law",
  "perse"= "adm_lic_revoc_law",
  "totfat" = "total_fatalities",
  "nghtfat" = "night_fatalities",
  "wkndfat" = "weekend_fatalities",
  "totfatpvm" = "total_fatal_per_100mm",
  "nghtfatpvm" = "night_fatal_per_100mm" ,
  "wkndfatpvm" = "weekend_fatal_per_100mm",
  "statepop" = "state_population",
  "totfatrte" = "total_fatality_rate",
  "nghtfatrte" = "night_fatality_rate",
  "wkndfatrte" = "weekend_fatality_rate",
  "vehicmiles" = "vehicle_miles_traveled_billions",
```

```
  "unem" = "unemployment_rate",
  "sbprim" = "primary_seatbelt_law",
  "sbsecon" = "secondary_seatbelt_law",
  "perc14_24" = "percent_pop_aged_14_to_24",
  "vehicmilespc" = "miles_driven_per_capita")
  )

# add a column for the 2 letter state code
states_2ltr_code <- sort(c(state.abb,"DC"))
state_code = c()
for (i in traffic_df$state) {
  state_code = c(state_code, states_2ltr_code[i])
}
traffic_df$state_code <- state_code

colnames(traffic_df)
```

```
##  [1] "year"                             "state"
##  [3] "sl55"                             "sl65"
##  [5] "sl70"                             "sl75"
##  [7] "slnone"                           "seatbelt"
##  [9] "minage"                           "zerotol"
## [11] "grad_drivers_lic_law"             "bac10"
## [13] "bac08"                            "adm_lic_revoc_law"
## [15] "total_fatalities"                 "night_fatalities"
## [17] "weekend_fatalities"               "total_fatal_per_100mm"
## [19] "night_fatal_per_100mm"            "weekend_fatal_per_100mm"
## [21] "state_population"                 "total_fatality_rate"
## [23] "night_fatality_rate"              "weekend_fatality_rate"
## [25] "vehicle_miles_traveled_billions" "unemployment_rate"
## [27] "percent_pop_aged_14_to_24"        "sl70plus"
## [29] "primary_seatbelt_law"             "secondary_seatbelt_law"
## [31] "d80"                              "d81"
## [33] "d82"                              "d83"
## [35] "d84"                              "d85"
## [37] "d86"                              "d87"
## [39] "d88"                              "d89"
## [41] "d90"                              "d91"
## [43] "d92"                              "d93"
## [45] "d94"                              "d95"
## [47] "d96"                              "d97"
## [49] "d98"                              "d99"
## [51] "d00"                              "d01"
## [53] "d02"                              "d03"
## [55] "d04"                              "miles_driven_per_capita"
## [57] "speed_limit"                      "year_of_observation"
## [59] "blood_alc_lim_10"                 "blood_alc_lim_08"
## [61] "state_code"
```

```
# convert data frame to pdata.frame
traffic_pdf <- pdata.frame(traffic_df, index=c("state", "year"))

## Check the structure of panel data
pdim(traffic_pdf)
```

```
## Balanced Panel: n = 48, T = 25, N = 1200
```

```
head(traffic_pdf, 10)
```

```
##        year state  sl55  sl65 sl70 sl75 slnone seatbelt minage zerotol
## 1-1980 1980     1 1.000 0.000    0    0      0        0     18       0
## 1-1981 1981     1 1.000 0.000    0    0      0        0     18       0
## 1-1982 1982     1 1.000 0.000    0    0      0        0     18       0
## 1-1983 1983     1 1.000 0.000    0    0      0        0     18       0
## 1-1984 1984     1 1.000 0.000    0    0      0        0     18       0
## 1-1985 1985     1 1.000 0.000    0    0      0        0     20       0
## 1-1986 1986     1 1.000 0.000    0    0      0        0     21       0
## 1-1987 1987     1 0.542 0.458    0    0      0        0     21       0
## 1-1988 1988     1 0.000 1.000    0    0      0        0     21       0
## 1-1989 1989     1 0.000 1.000    0    0      0        0     21       0
##        grad_drivers_lic_law bac10 bac08 adm_lic_revoc_law total_fatalities
## 1-1980                    0     1     0                 0              940
## 1-1981                    0     1     0                 0              933
## 1-1982                    0     1     0                 0              839
## 1-1983                    0     1     0                 0              930
## 1-1984                    0     1     0                 0              932
## 1-1985                    0     1     0                 0              882
## 1-1986                    0     1     0                 0             1080
## 1-1987                    0     1     0                 0             1111
## 1-1988                    0     1     0                 0             1024
## 1-1989                    0     1     0                 0             1029
##        night_fatalities weekend_fatalities total_fatal_per_100mm
## 1-1980              422                236                 3.200
## 1-1981              434                248                 3.350
## 1-1982              376                224                 2.810
## 1-1983              397                223                 3.000
## 1-1984              421                237                 2.830
## 1-1985              358                224                 2.510
## 1-1986              500                279                 3.177
## 1-1987              499                300                 2.970
## 1-1988              423                226                 2.580
## 1-1989              418                247                 2.520
##        night_fatal_per_100mm weekend_fatal_per_100mm state_population
## 1-1980                 1.437                  0.803          3893888
## 1-1981                 1.558                  0.890          3918520
## 1-1982                 1.259                  0.750          3925218
## 1-1983                 1.281                  0.719          3934109
## 1-1984                 1.278                  0.720          3951834
## 1-1985                 1.019                  0.637          3972527
## 1-1986                 1.471                  0.821          3991569
## 1-1987                 1.334                  0.802          4015261
## 1-1988                 1.066                  0.569          4023858
## 1-1989                 1.024                  0.605          4030229
##        total_fatality_rate night_fatality_rate weekend_fatality_rate
## 1-1980               24.14                10.84                 6.06
## 1-1981               24.07                11.08                 6.33
## 1-1982               21.37                 9.58                 5.71
## 1-1983               23.64                10.09                 5.67
## 1-1984               23.58                10.65                 6.00
## 1-1985               22.20                 9.01                 5.64
```

```
## 1-1986              27.08              12.53                  6.99
## 1-1987              27.67              12.43                  7.47
## 1-1988              25.45              10.51                  5.62
## 1-1989              25.53              10.37                  6.13
##        vehicle_miles_traveled_billions unemployment_rate
## 1-1980                        29.37500              8.8
## 1-1981                        27.85200             10.7
## 1-1982                        29.85765             14.4
## 1-1983                        31.00000             13.7
## 1-1984                        32.93286             11.1
## 1-1985                        35.13944              8.9
## 1-1986                        33.99371              9.8
## 1-1987                        37.40741              7.8
## 1-1988                        39.68992              7.2
## 1-1989                        40.83333              7.0
##        percent_pop_aged_14_to_24 sl70plus primary_seatbelt_law
## 1-1980                      18.9        0                    0
## 1-1981                      18.7        0                    0
## 1-1982                      18.4        0                    0
## 1-1983                      18.0        0                    0
## 1-1984                      17.6        0                    0
## 1-1985                      17.3        0                    0
## 1-1986                      17.0        0                    0
## 1-1987                      16.6        0                    0
## 1-1988                      16.2        0                    0
## 1-1989                      15.8        0                    0
##        secondary_seatbelt_law d80 d81 d82 d83 d84 d85 d86 d87 d88 d89 d90 d91
## 1-1980                      0   1   0   0   0   0   0   0   0   0   0   0   0
## 1-1981                      0   0   1   0   0   0   0   0   0   0   0   0   0
## 1-1982                      0   0   0   1   0   0   0   0   0   0   0   0   0
## 1-1983                      0   0   0   0   1   0   0   0   0   0   0   0   0
## 1-1984                      0   0   0   0   0   1   0   0   0   0   0   0   0
## 1-1985                      0   0   0   0   0   0   1   0   0   0   0   0   0
## 1-1986                      0   0   0   0   0   0   0   1   0   0   0   0   0
## 1-1987                      0   0   0   0   0   0   0   0   1   0   0   0   0
## 1-1988                      0   0   0   0   0   0   0   0   0   1   0   0   0
## 1-1989                      0   0   0   0   0   0   0   0   0   0   1   0   0
##        d92 d93 d94 d95 d96 d97 d98 d99 d00 d01 d02 d03 d04
## 1-1980   0   0   0   0   0   0   0   0   0   0   0   0   0
## 1-1981   0   0   0   0   0   0   0   0   0   0   0   0   0
## 1-1982   0   0   0   0   0   0   0   0   0   0   0   0   0
## 1-1983   0   0   0   0   0   0   0   0   0   0   0   0   0
## 1-1984   0   0   0   0   0   0   0   0   0   0   0   0   0
## 1-1985   0   0   0   0   0   0   0   0   0   0   0   0   0
## 1-1986   0   0   0   0   0   0   0   0   0   0   0   0   0
## 1-1987   0   0   0   0   0   0   0   0   0   0   0   0   0
## 1-1988   0   0   0   0   0   0   0   0   0   0   0   0   0
## 1-1989   0   0   0   0   0   0   0   0   0   0   0   0   0
##        miles_driven_per_capita speed_limit year_of_observation blood_alc_lim_10
## 1-1980                7543.874          55                1980                1
## 1-1981                7107.785          55                1981                1
## 1-1982                7606.622          55                1982                1
## 1-1983                7879.802          55                1983                1
## 1-1984                8333.562          55                1984                1
```

```
## 1-1985               8845.614          55               1985               1
## 1-1986               8516.377          55               1986               1
## 1-1987               9316.308          55               1987               1
## 1-1988               9863.649          NA               1988               1
## 1-1989              10131.764          NA               1989               1
##        blood_alc_lim_08 state_code
## 1-1980                0         AK
## 1-1981                0         AK
## 1-1982                0         AK
## 1-1983                0         AK
## 1-1984                0         AK
## 1-1985                0         AK
## 1-1986                0         AK
## 1-1987                0         AK
## 1-1988                0         AK
## 1-1989                0         AK
```

2. (5 points) Provide a description of the basic structure of the dataset. What is this data? How, where, and when is it collected? Is the data generated through a survey or some other method? Is the data that is presented a sample from the population, or is it a *census* that represents the entire population? Minimally, this should include:
   - How is the our dependent variable of interest `total_fatalities_rate` defined?

The data set that we plan to analyze is what is called Panel Data. The data set contains observations for every state (in the continental USA, which is made of 48 states in the contiguous land, except Delaware, and including Washington DC) from 1980 to 2004 (25 years) of vehicular accident data.

The data is organized as 25 rows (one for each calendar year) for every state. The total number of rows is 1200 **which is 48 states times 25 years. The sequence for each states is from 1980 to 2004. All the data for a given state is listed for all 25 years before the next state is taken up. The states are organized in alphabetical order.**

2.1 Fatalities Information

Primarily, for each year and for every state the data set measures various fatalities (total fatalities, night fatalities, weekend fatalities, and their respective rates by for every 100,000 people in the state and per every 100 Million Mile driven). The data contains several driving laws for each state (such as speed limits, blood alcohol content level for declaring driver being intoxicated, minimum driving age, zero-tolerence policy etc.). There is one column "perse" in the original data set (that we have renamed it as "adm_lic_revoc_law"), that tells if the state can revoke license without a trial, and seat belt laws. There are also demographic details (population age) and economic data (unemployment rate).

A few state law variables, such as bac08 and bac10 are coded as dichotomous options, i.e. 1 and 0, indicating whether the law was applicable in that state in that year. A fraction value indicates that fraction of the year for which that law was applicable in that state. e.g. a fraction value of 0.667 means that the law was implemented in that state for 2/3 rd of the duration in that year.

2.2 Demographic Information

'vehicle_miles_traveled_billions', 'miles_driven_per_capita', 'unemployment_rate' and 'percent_pop_aged_14_to_24' are the four demographic variables. 'vehicle_miles_traveled_billions' and 'miles_driven_per_capita' account for overall miles driven and per capita miles driven by state's population. 'unemployment_rate'

represents the rate of unemployment in a state in a given year. 'percent_pop_aged_14_to_24' represents how much % of state's population is youth.

2.3 Traffic Laws

There are quite a few variables explaining traffic laws across states over time.

Seatbelt related laws, i.e. 'primary_seatbelt_law' and 'secondary_seatbelt_law' are categorical variables with '0' indicating no law, '1' indicating primary law (no other violation needed to issue ticket) and '2' indicating secondary law (There must be at least 1 other violation to issue ticket)

Drunk Driviing related laws, i.e. 'minage', 'zerotol', 'adm_lic_revoc_law', 'blood_alc_lim_08' and 'blood_alc_lim_10' are the laws to prevent drunk driving in a state. 'minage' reflects minimum age required to issue a driver's license. The range varies from 18 - 21 years. 'zerotol' indicates DUI offense charge for any drunk driving related traffic violation. The range varies between 0 and 1. 'adm_lic_revoc_law' enforces revocation of driver's license for drunk driver related offenses. A fraction reflects portion of the year for which the law was implemented. 'blood_alc_lim_08' and 'blood_alc_lim_10' reflect whether the allowed blood alcohol level in a driver to avoid offense is 0.08% or 0.10% respectively.

Speed Limit related laws, i.e. 'sl55', 'sl60', 'sl65', 'sl70', 'sl75' and 'sl70plus' are dummy variables indicating implemented speed limit in a state in a given year.

2.4 Data Source

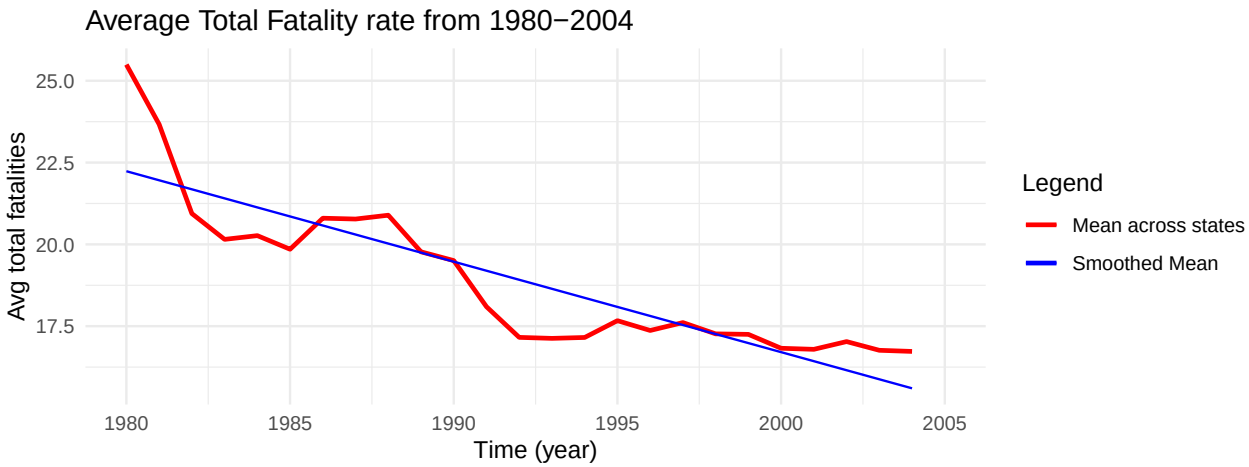
Very likely, the data is collected by Insurance Institute for Highway Safety (IIHS) or The National Highway Traffic Safety Administration (NHTSA) aided by reporting from local law enforcement. Local law enforcement, such as county police, highway patrol, or other agencies attend to most of the incidents, and report on these incidents. These are then collected, curated, and maintained by agencies such as NHTSA, IIHS for analysis and for policy formulation aimed toward reducing accidents and fatalities. The data is summarizes at yearly level for each state based on individual reports from state/county/town law enforcement, high patrol, or other similar agencies. This data doesn't appear to have been generated by a survey. It will be a non-trivial task for a survey respondent to estimate weekend fatalities, night fatalities etc., and more so the respective rates by population and/or vehicle miles driven.

The data appears to have been census driven across the state population, and then summarized. This data is not a sample from the population.

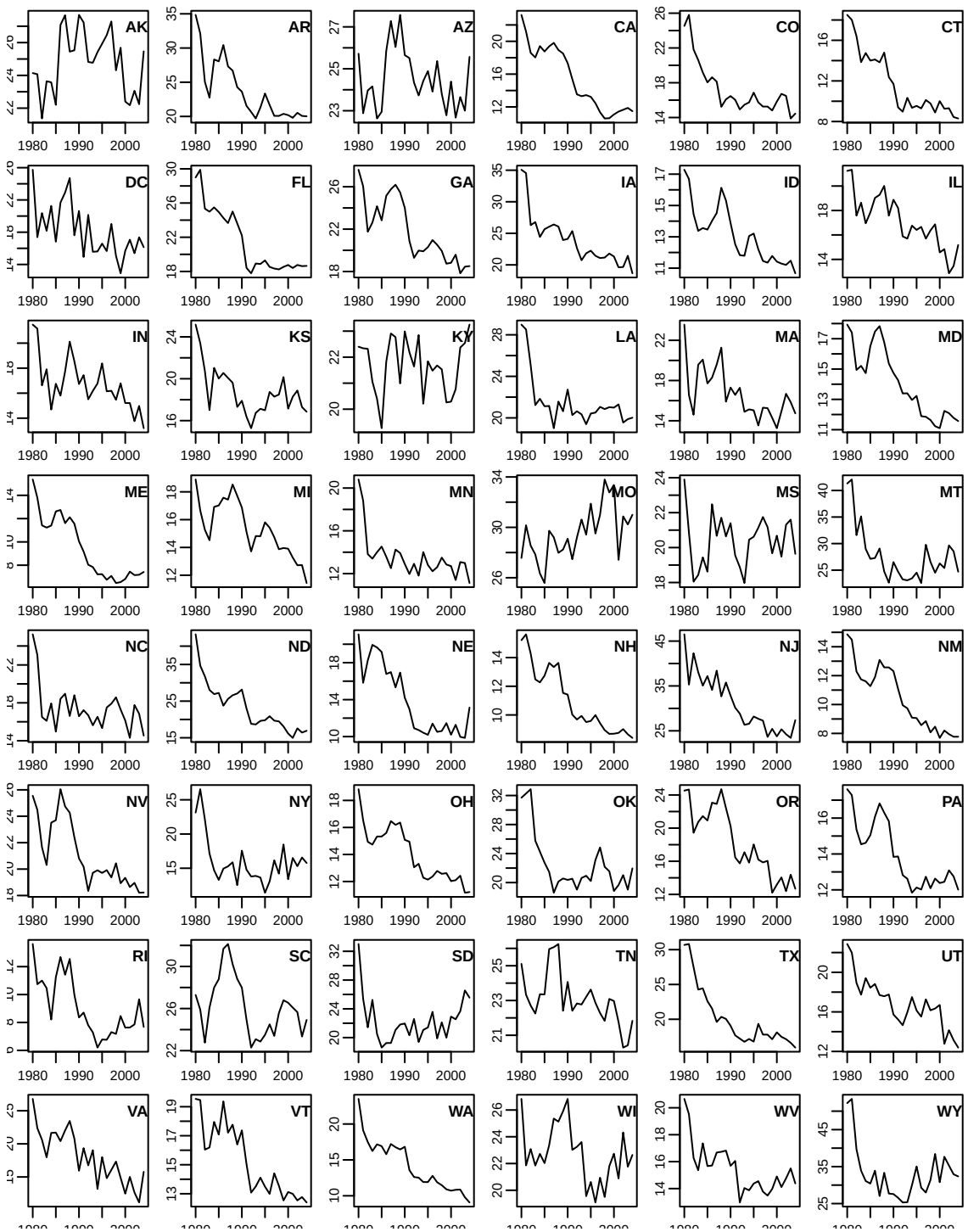
The variable "totfatrate" in the original data, which we have renamed as "total_fatality_rate" is the total number of fatal accidents per 100,000 people (all of whom may not be drivers). In the data set we also see weekend and night fatalities metrics. The rest (after subtracting night and weekend fatalities) gives the fatal accidents number during day time on weekdays.

3. (20 points) Conduct a very thorough EDA, which should include both graphical and tabular techniques, on the dataset, including both the dependent variable `total_fatalities_rate` and the potential explanatory variables. Minimally, this should include:
   - How is the our dependent variable of interest `total_fatalities_rate` defined?
   - What is the average of `total_fatalities_rate` in each of the years in the time period covered in this dataset?

As with every EDA this semester, the goal of this EDA is not to document your own process of discovery – save that for an exploration notebook – but instead it is to bring a reader that is new to the data to a full understanding of the important features of your data as quickly as possible. In order to do this, your EDA should include a detailed, orderly narrative description of what you want your reader to know. Do not include any output – tables, plots, or statistics – that you do not intend to write about.



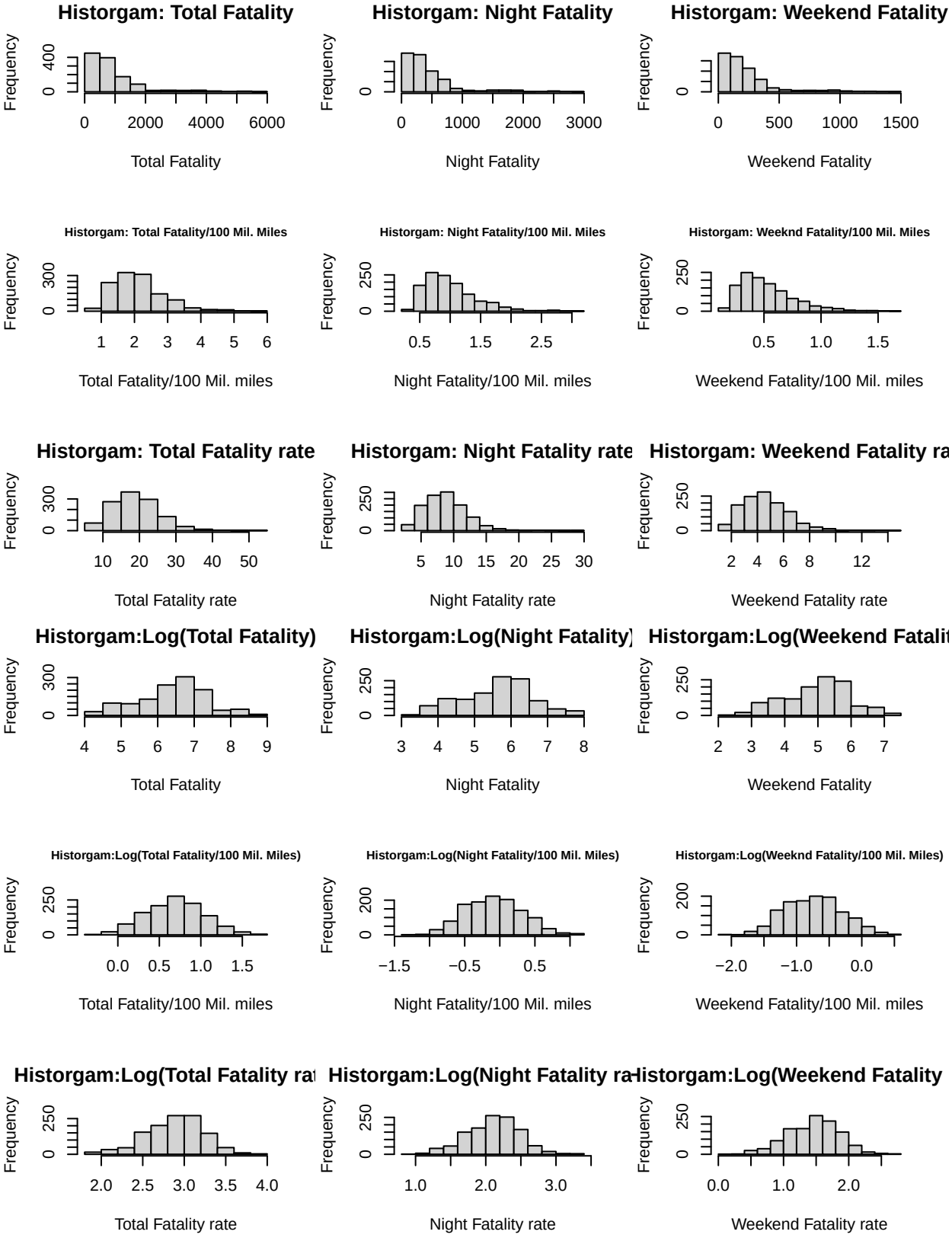
The plot above shows the average across states for a given year. We show both the average value as well as the smoothed value. There is definitely a downward trend in the number of accidents.



The plots above capture the traffic fatality rate (per 100,000 population) for each state. While most states show a general downward trend in the number of fatal rate there are some exceptions

- AK (Arkansas), AZ (Arizona), and MO (Missouri): These two states show upward trends before dipping down toward the very end.
- KY (Kentucky) and MS (Mississippi): This state shows a jagged saw-tooth pattern.

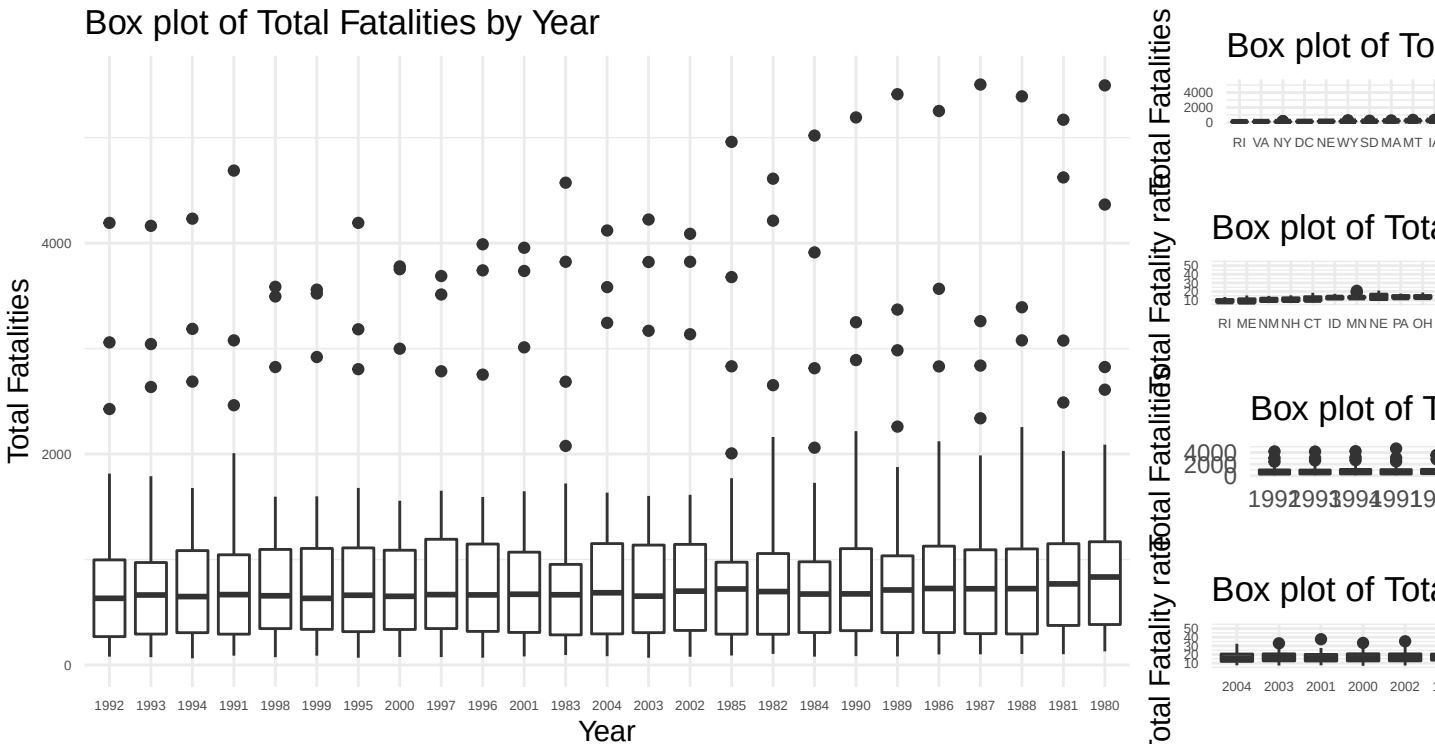
- SD (South Dakota): There is an initial steep decrease followed by moderate upward trend.



The two set of plots above show the histogram of the outcome variables (various fatality

measure) in its native form and with log transformation. The log-transformed data shows a behavior that is more close to that of a normal distribution. As part of the modeling analysis we also conducts formal test (using Shapiro-Wilk's method). We take advantage of the log transformation in the modeling and analysis.

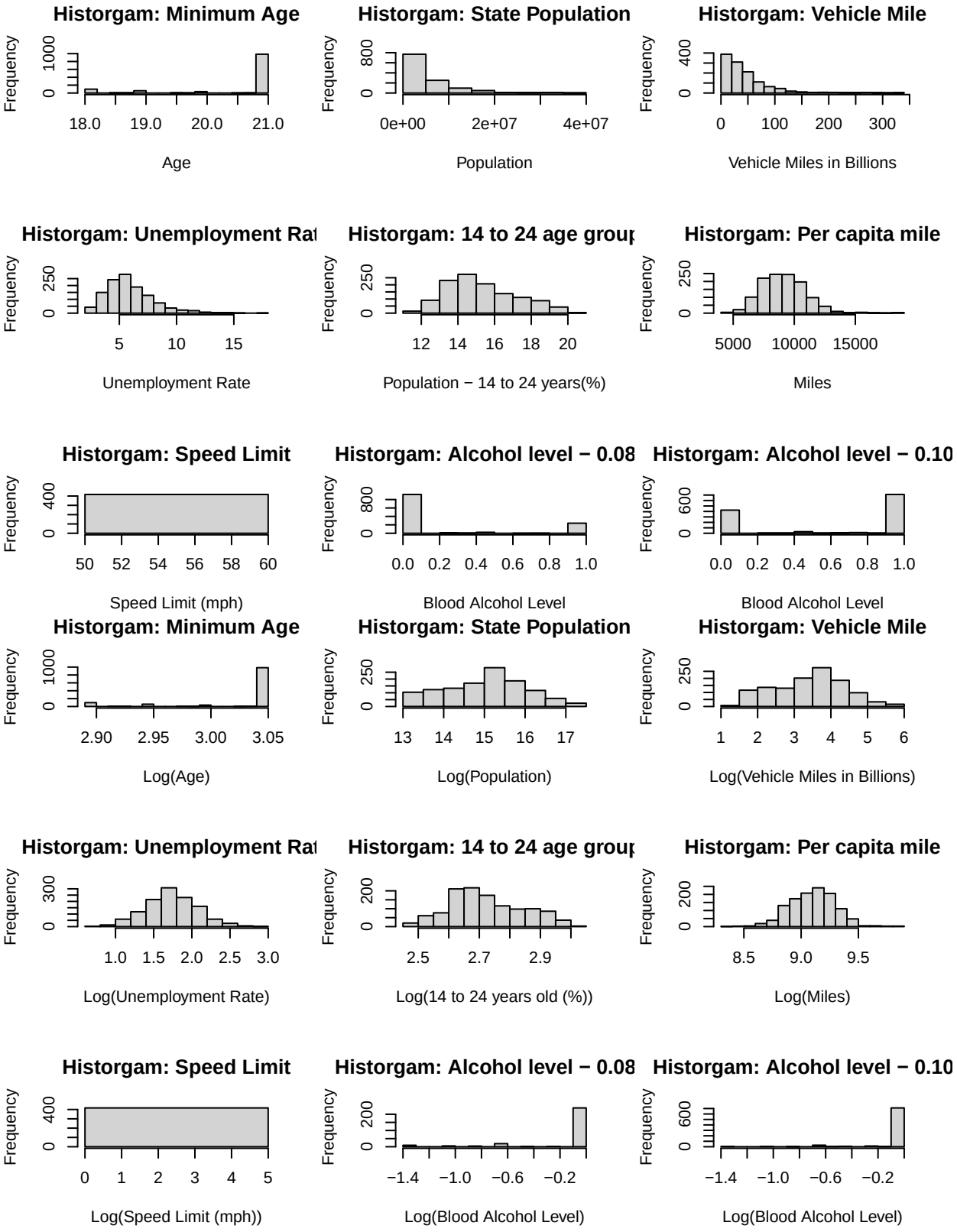
In the graphs below we show the box plots of the total fatality and total fatality rates by state and by year. Unlike the histogram, where we have shown all the dependent variables (such as nigh/weekend metrics), we focus on the total fatalities related outcomes, both by state and year.



The box plots reveal what is generally to be expected. We see a higher number with California (CA) state, for total fatality, due to the high population and generally being a somewhat of a larger state where people need to drive moderate distances, outside of major cities like Los Angeles, San Francisco, and a few others. However, if we consider fatality rate, we see that states like New Jersey (NJ), Missouri (MO), and Wyoming (WY) are leading the pack.

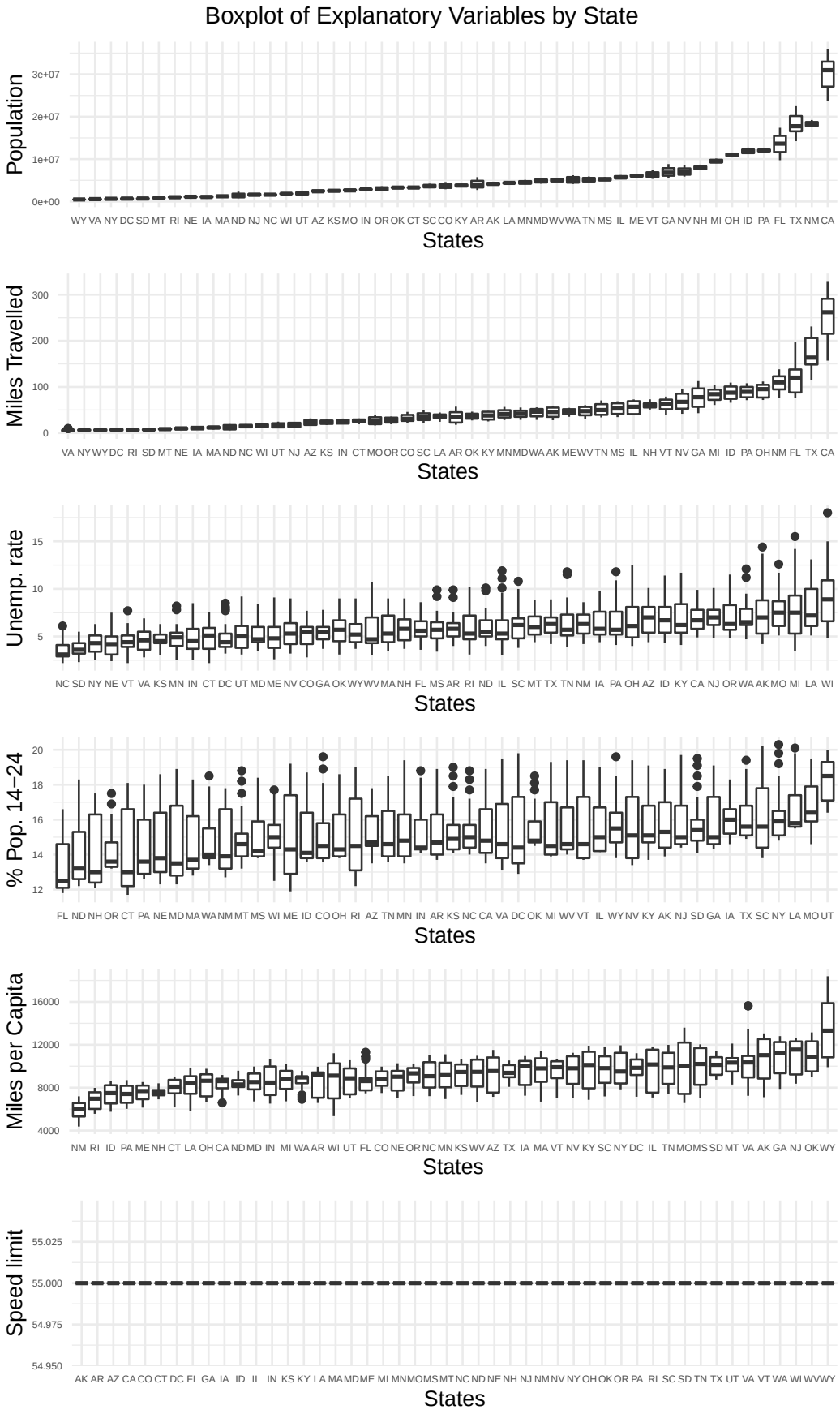
When we look at the next two box plots, which show data across states by year, we see quite a bit of outliers. This is primarily due to the varying conditions of population, driving distances, and other demographic/economic factors. For instance, if we consider California and Wyoming, the conditions are vastly different on demographics, economics, laws such as speed limits, distance driven etc. Thus, the yearly plot tend to bring out the outliers among states. As an additional observation, we see more of these outliers when we look at the absolute fatality number, and less so with fatality rate.

We will now examine the histogram and box plots of the explanatory variables



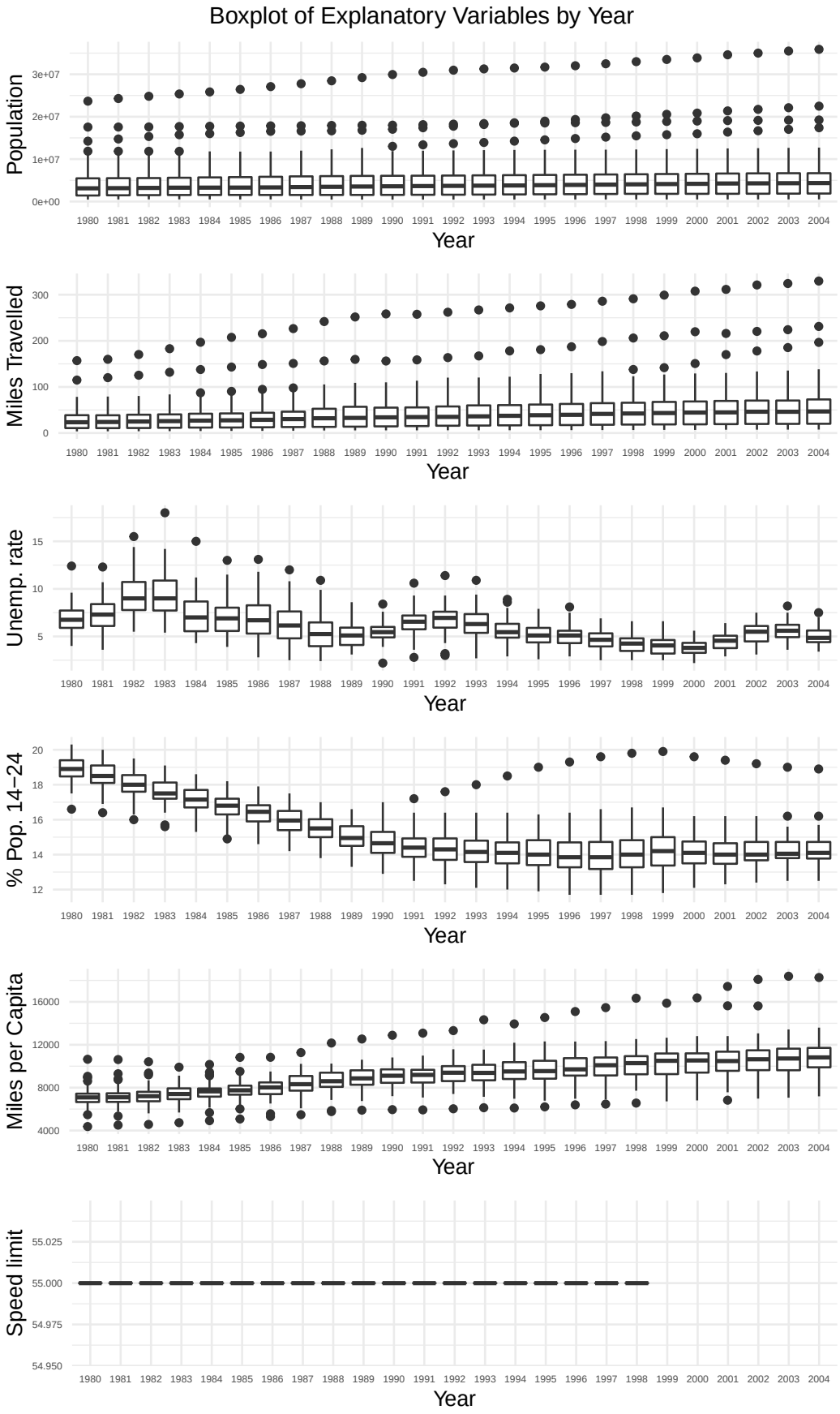
From the histograms above we see that some of the left-skewed histogram of the variables in its natural form is coming closer to that of a normal distribution, when log-transformed. Notable exceptions to the statement just made are the discrete variables, in particular, Minimum Age

and Speed Limit variables. The reasons are obvious. Log transformation doesn't change the fundamentally non-continuous nature of these variables. As alluded to earlier, we will take advantage of the log-transform in the models.


Boxplot of Explanatory Variables by State
Population
3e+07
2e+07
1e+07
0e+00
WY VA NY DC SD MT RI NE IA MA ND NJ NC WI UT AZ KS MO IN OR OK CT SC CO KY AR AK LA MN MD WV WA TN MS IL ME VT GA NV NH MI OH ID PA FL TX NM CA
States
Miles Travelled
300
200
100
0
VA NY WY DC RI SD MT NE IA MA ND NC WI UT NJ AZ KS IN CT MO OR CO SC LA AR OK KY MN MD WA AK ME WV TN MS IL NH VT NV GA MI ID PA OH NM FL TX CA
States
Unemp. rate
15
10
5
NC SD NY NE VT VA KS MN IN CT DC UT MD ME NV CO GA OK WY WV MA NH FL MS AR RI ND IL SC MT TX TN NM IA PA OH AZ ID KY CA NJ OR WA AK MO MI LA WI
States
% Pop. 14–24
20
18
16
14
12
FL ND NH OR CT PA NE MD MA WA NM MT MS WI ME ID CO OH RI AZ TN MN IN AR KS NC CA VA DC OK MI WV VT IL WY NV KY AK NJ SD GA IA TX SC NY LA MO UT
States
Miles per Capita
16000
12000
8000
4000
NM RI ID PA ME NH CT LA OH CA ND MD IN MI WA AR WI UT FL CO NE OR NC MN KS WV AZ TX IA MA VT NV KY SC NY DC IL TN MO MS SD MT VA AK GA NJ OK WY
States
Speed limit
55.025
55.000
54.975
54.950
AK AR AZ CA CO CT DC FL GA IA ID IL IN KS KY LA MA MD ME MI MN MO MS MT NC ND NE NH NJ NM NV NY OH OK OR PA RI SC SD TN TX UT VA VT WA WI WV WY
States

We see from the box plots above, of the explanatory variables by state, CA has the highest population, mean unemployment rate is very close among the states with a few outliers, and the miles traveled is led by CA, being a populous and a large state. However, the miles per-capita is being led by Wyoming. This is not entirely unexpected. A state as vast a Wyoming and with a smaller population people do drive a lot in Wyoming to get around. It is also a rural state where people don't find a lot of what they want nearby. Additionally, it's not uncommon to see animals and produce transported in smaller trucks from farms to markets. As for speed limits, it appears that that are just two average speed limits across states - 55mph and 65mph. We do see lower speed limits in Montana (MT). We also see higher speed limits of 70mph is some states. As highway quality and vehicle features improve we see higher speed limits.

```
## Warning: Removed 783 rows containing non-finite values (stat_boxplot).
```

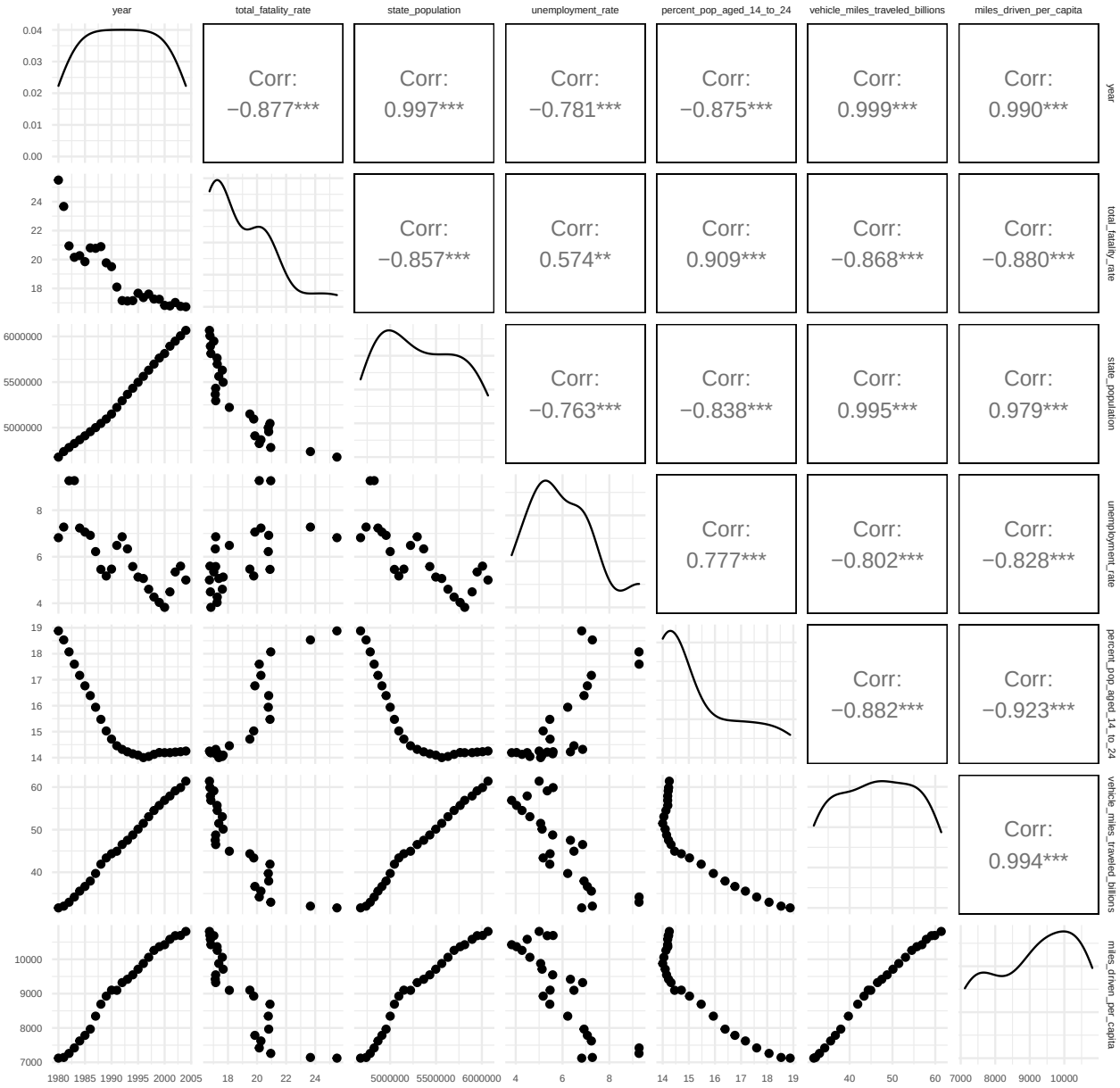

Boxplot of Explanatory Variables by Year
Population
3e+07
2e+07
1e+07
0e+00
Miles Travelled
300
200
100
0
Unemp. rate
15
10
5
% Pop. 14–24
20
18
16
14
12
Miles per Capita
16000
12000
8000
4000
Speed limit
55.025
55.000
54.975
54.950
1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004
Year

The box plots above are across states for each year. The population averages across states has remained more or less same across years. We see a lot of outliers due to movement of people. Relocation happens every so often for various reasons, and there may be an instantaneous increase or decrease of population in some states in some of the years. The outliers reflect this reality. The miles traveled per-capita shows a steady increase albeit by a small amount. The speed limit has gone up steadily. As we make better roads and vehicles, and improve fuel efficiency, people tend to drive more. Areas that didn't have roads have opened up leading to residential buildings in places that were once off-limit. This has caused more driving. The average unemployment rate has remined nearly constant barring blips in 1981-82, 1991-92, and 2001-2002; the classic 10-year cycle. The population in 14-24 age group decresed initially, but seems to be going up although very moderately. The outliers capture the uneven distribution across states.

```
# annual mean
annual_mean <- aggregate(
  traffic_df[, c("total_fatality_rate", "state_population",
                 "unemployment_rate","percent_pop_aged_14_to_24",
                 "vehicle_miles_traveled_billions", "miles_driven_per_capita")],
  traffic_df["year"], FUN = mean)

ggpairs(annual_mean, size = 5, cardinality_threshold = 25) +
  theme(strip.text.x = element_text(size = 5),
        strip.text.y = element_text(size = 5)) +
  theme(axis.text = element_text(size = 5)) +
  ggtitle(label = "State Averages over time (years)")
```

State Averages over time (years)

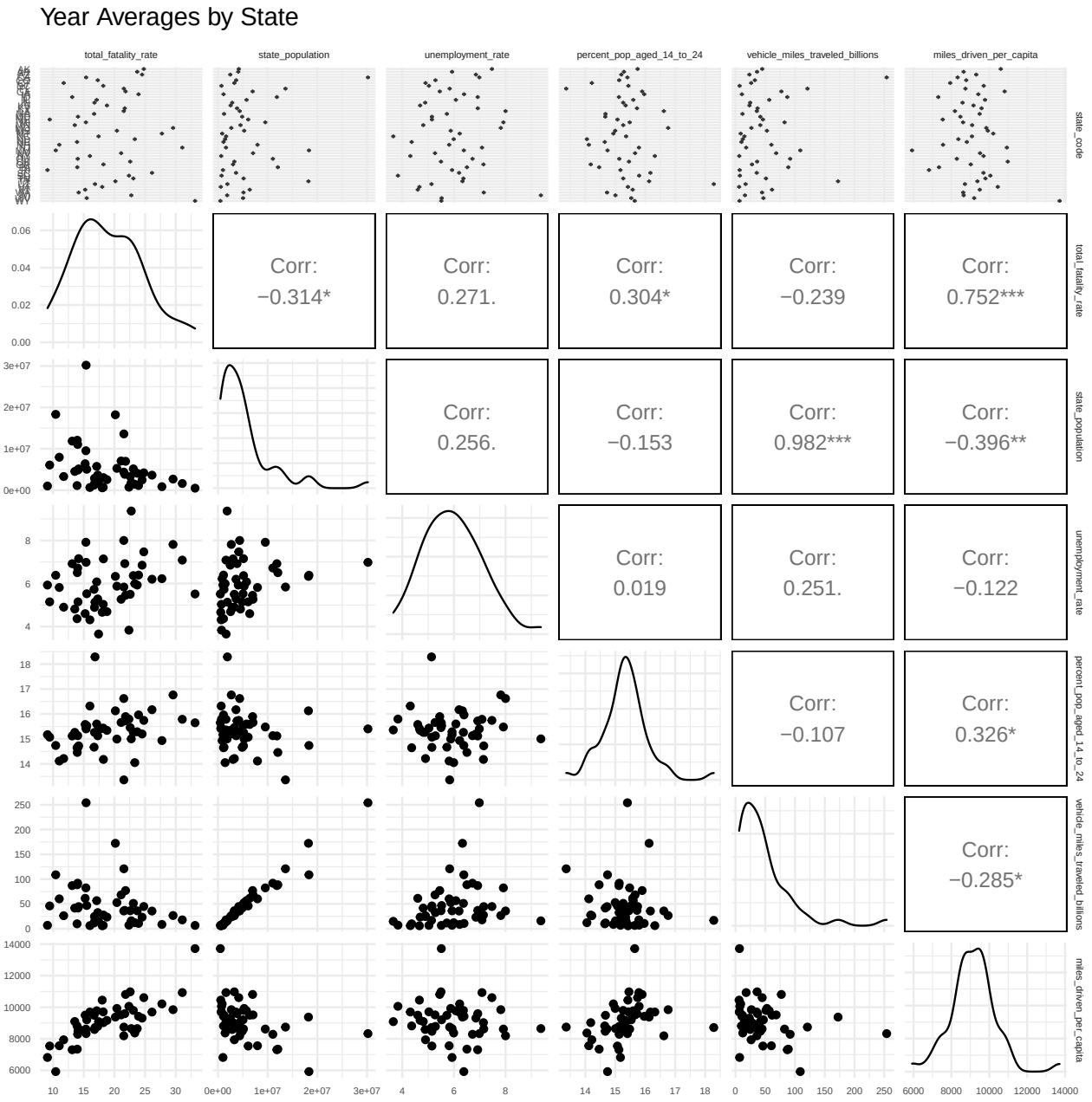



```
# state mean
state_mean <- aggregate(
  traffic_df[, c("total_fatality_rate", "state_population",
                 "unemployment_rate", "percent_pop_aged_14_to_24",
                 "vehicle_miles_traveled_billions","miles_driven_per_capita")],
  traffic_df["state_code"], FUN = mean)

s <- ggpairs(state_mean, size = 5, cardinality_threshold = 60) +
       theme(strip.text.x = element_text(size = 5),
             strip.text.y = element_text(size = 5)) +
       theme(axis.text = element_text(size = 5)) +
       ggtitle(label = "Year Averages by State")

s$plots <- s$plots[-(seq(1, s$ncol*s$nrow, by = s$ncol))]
```

```
s$ncol <- s$ncol - 1
s$xAxisLabels <- s$xAxisLabels[-1]
s
```



The first plot above show the cross-correlations of among annual mean of total fatality rate against economic, demographic, and vehicle miles.

Observations: 1. *state_population* increase with time which makes sense 2. *vehicle_miles_traveled_billions* increase with time as more people drive more miles over years 3. *miles_driven_per_capita* increase with time probably because cars are more affordable and hence accessible 4. *percent_pop_aged_14_to_24* decreases indicating aged and experienced population resulting into fewer fatalities 5. *unemployment_rate* is cyclical but decreasing overall over years 6. *total_fatality_rate* decreases with time as we observed earlier

We observe covariance between pairs of variables as well but that probably is more of a spurious correlation.

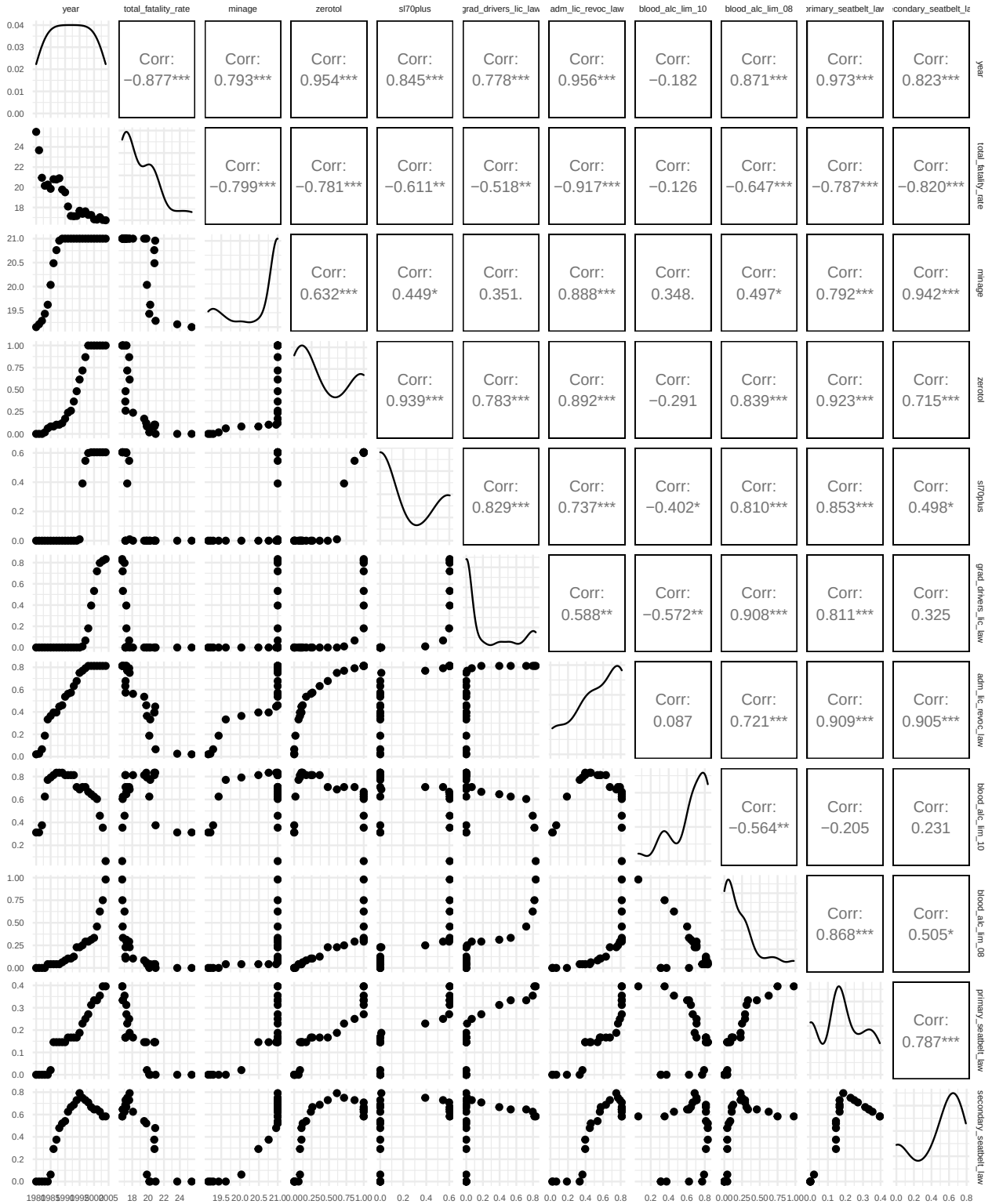
We also show the mean for each state of fatality rate and its cross-correlation with the same set of variables. Observations: 1. *total_fatality_rate* shows high positive correlation with *miles_driven_per_capita* meaning more time spent on road means increased probability of fatality. 2. *total_fatality_rate* shows moderate positive correlation with *percent_pop_aged_14_to_24* meaning higher youth proportion, higher their fatality rate. 3. *percent_pop_aged_14_to_24* shows moderate positive correlation with *miles_driven_per_capita* meaning states with higher young generation drives more, resulting into higher fatality rate, which could be due to negligence driving habits of younger drivers overall. 4. *state_population* shows strong positive correlation with *vehicle_miles_traveled_billions* meaning more people drive more. 5. *state_population* shows moderate negative correlation with *total_fatality_rate* meaning more people more enforcing safe driving infrastructure. 6. *state_population* shows moderate negative correlation with *miles_driven_per_capita* meaning states with larger population drive lesser, leading to lower on road fatalities

We now show similar cross-correlation for the total fatality rate with, both annually and by state, on the laws of the states

```
annual_law_mean <- aggregate(
  traffic_df[,c("total_fatality_rate", "minage", "zerotol", "sl70plus",
                "grad_drivers_lic_law", "adm_lic_revoc_law", "blood_alc_lim_10",
                "blood_alc_lim_08", "primary_seatbelt_law",
                "secondary_seatbelt_law")],
  traffic_df["year"], FUN = mean)

ggpairs(annual_law_mean, size = 5, cardinality_threshold=100,
        upper = list(continuous = wrap("cor", size = 3))) +
  theme(strip.text.x = element_text(size = 5),
        strip.text.y = element_text(size = 5)) +
  theme(axis.text = element_text(size = 5)) +
  ggtitle(label = "State Law Proportions over time (years)")
```

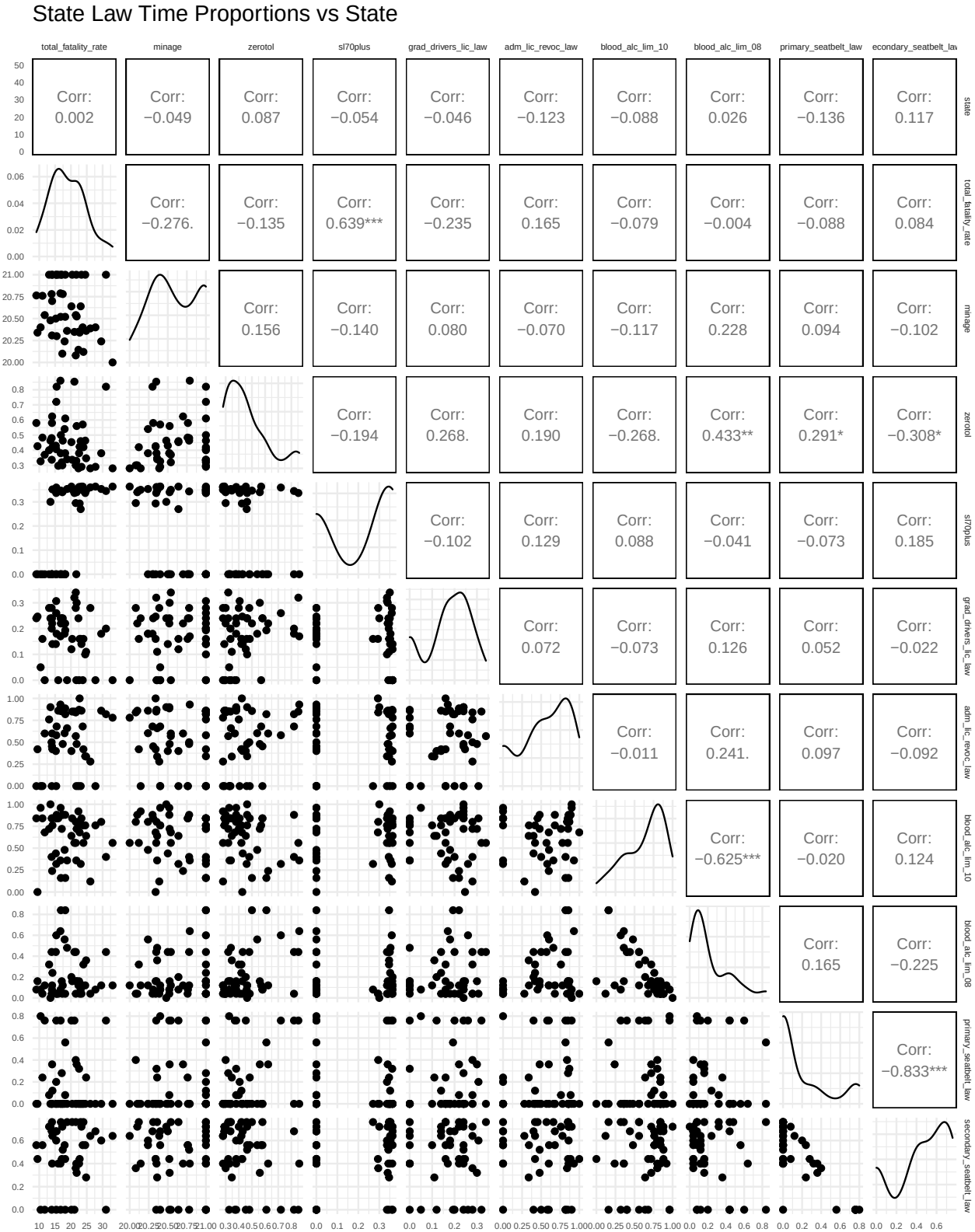
State Law Proportions over time (years)




```
state_law_mean <- aggregate(
  traffic_df[,c("total_fatality_rate", "minage", "zerotol", "sl70plus",
                "grad_drivers_lic_law", "adm_lic_revoc_law", "blood_alc_lim_10",
                "blood_alc_lim_08", "primary_seatbelt_law",
```

```
                "secondary_seatbelt_law")],
  traffic_df["state"], FUN = mean)
l <- ggpairs(state_law_mean, size = 5, cardinality_threshold=48,
             upper = list(continuous = wrap("cor", size = 3))) +
      theme(strip.text.x = element_text(size = 5),
             strip.text.y = element_text(size = 5)) +
     theme(axis.text = element_text(size = 5)) +
     ggtitle(label = "State Law Time Proportions vs State")

l$plots <- l$plots[-(seq(1, l$ncol*l$nrow, by = l$ncol))]
l$ncol <- l$ncol - 1
l$xAxisLabels <- l$xAxisLabels[-1]
l
```



The first plot above show the cross-correlations of proportions of time (years) states implemented various traffic law verses, i.e. how the adaptability of these laws have changed over time.

Observations: 1. *grad_drivers_lic_law* shows strong positive correlation with *primary_seatbelt_law* meaning if in any given year, a large number of states has implemented one law, then large number of states will implement the other law as well 2. *blood_alc_lim_08* shows high positive correlation over time indicating that larger and larger number of states have implemented that law with time. 3. *blood_alc_lim_10* shows low negative correlation over time indicating that lesser number of states have implemented that law with time. It is expected given that *blood_alc_lim_08* is more popular among states 4. Other laws *minage, zerotol, sl70plus, grad_drivers_lic_law, adm_lic_revoc_law* also show string positive correlation over time indicating that larger and larger number of states have implemented that law with time.

The second plot above show the cross-correlations of proportions of time (years) states implemented various traffic law verses states

Observations: 1. *primary_seatbelt_law* shows string negative correlation with *secondary_seatbelt_law* meaning if a state has implemented one law, they dont implement other and it makes sense as these laws appear mutually exclusive 2. *blood_alc_lim_08* shows high negtive correlation with *blood_alc_lim_10* meaning state would implement one of the two. In other words, if a state implements one law for larger number of years, it would not implements the other. 3. *zerotol* shows moderate positive correlation with *blood_alc_lim_08* meaning longer one state one of the law implemented, longer we expect that state to implement the other one two.

3 (15 points) Preliminary Model

Estimate a linear regression model of *totfatrte* on a set of dummy variables for the years 1981 through 2004 and interpret what you observe. In this section, you should address the following tasks:

- Why is fitting a linear model a sensible starting place?
- What does this model explain, and what do you find in this model?
- Did driving become safer over this period? Please provide a detailed explanation.
- What, if any, are the limitation of this model. In answering this, please consider **at least**:
  - Are the parameter estimates reliable, unbiased estimates of the truth? Or, are they biased due to the way that the data is structured?
  - Are the uncertainty estimate reliable, unbiased estimates of sampling based variability? Or, are they biased due to the way that the data is structured?

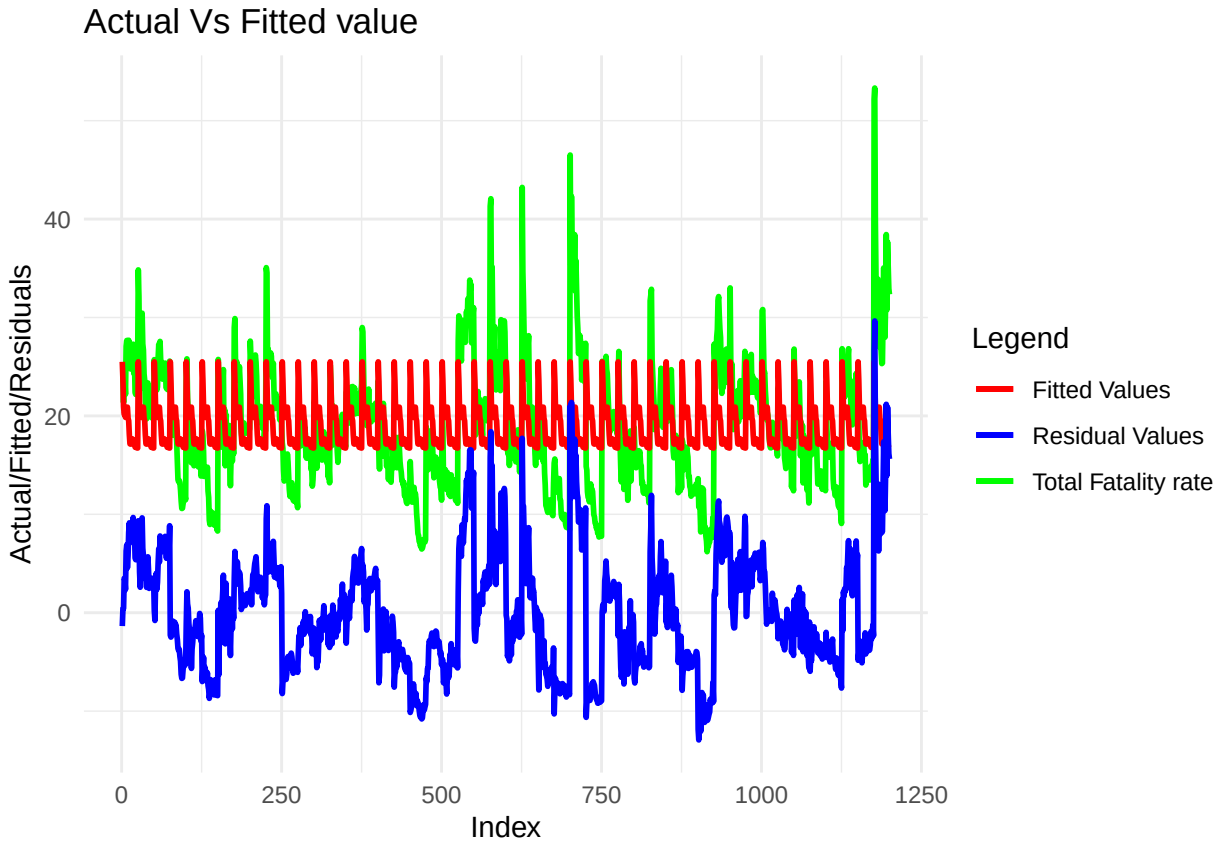
```
base_model <- lm(total_fatality_rate ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 +
                   d88 + d89 + d90 + d91 + d92 + d93 + d94 + d95 + d96 + d97 +
                   d98 + d99 + d00 + d01 + d02 + d03 + d04, data = traffic_df)
summary(base_model)
```

```
##
## Call:
## lm(formula = total_fatality_rate ~ d81 + d82 + d83 + d84 + d85 +
##     d86 + d87 + d88 + d89 + d90 + d91 + d92 + d93 + d94 + d95 +
##     d96 + d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04, data = traffic_df)
##
## 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 ***
## d81          -1.8244     1.2263  -1.488 0.137094
## d82          -4.5521     1.2263  -3.712 0.000215 ***
## d83          -5.3417     1.2263  -4.356 1.44e-05 ***
## d84          -5.2271     1.2263  -4.263 2.18e-05 ***
```

```
## d85          -5.6431     1.2263  -4.602 4.64e-06 ***
## d86          -4.6942     1.2263  -3.828 0.000136 ***
## d87          -4.7198     1.2263  -3.849 0.000125 ***
## d88          -4.6029     1.2263  -3.754 0.000183 ***
## d89          -5.7223     1.2263  -4.666 3.42e-06 ***
## d90          -5.9894     1.2263  -4.884 1.18e-06 ***
## d91          -7.3998     1.2263  -6.034 2.14e-09 ***
## d92          -8.3367     1.2263  -6.798 1.68e-11 ***
## d93          -8.3669     1.2263  -6.823 1.43e-11 ***
## d94          -8.3394     1.2263  -6.800 1.66e-11 ***
## d95          -7.8260     1.2263  -6.382 2.51e-10 ***
## d96          -8.1252     1.2263  -6.626 5.25e-11 ***
## d97          -7.8840     1.2263  -6.429 1.86e-10 ***
## d98          -8.2292     1.2263  -6.711 3.01e-11 ***
## d99          -8.2442     1.2263  -6.723 2.77e-11 ***
## d00          -8.6690     1.2263  -7.069 2.67e-12 ***
## d01          -8.7019     1.2263  -7.096 2.21e-12 ***
## d02          -8.4650     1.2263  -6.903 8.32e-12 ***
## d03          -8.7310     1.2263  -7.120 1.88e-12 ***
## d04          -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
```

```
aug_base_model <- augment(base_model)

aug_base_model %>%
  ggplot(aes(x=1:1200)) +
  geom_line(aes(y = total_fatality_rate, colour = "Total Fatality rate"),
            size = 1) +
  geom_line(aes(y = .fitted, colour = "Fitted Values"), size = 1) +
  geom_line(aes(y = .resid, colour = "Residual Values"), size = 1) +
  labs(title = "Actual Vs Fitted value", x = "Index",
       y = "Actual/Fitted/Residuals") +
  scale_colour_manual(name="Legend", values=c("red", "blue", "green"))
```




```
shapiro.test(aug_base_model$.resid) # test for normal distribution
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aug_base_model$.resid
## W = 0.9703, p-value = 5.637e-15
```

```
t.test(augment(base_model)$.fitted,
       augment(lm(total_fatality_rate ~ 1, data=traffic_df))$.fitted)
```

```
##
##  Welch Two Sample t-test
##
## data:  augment(base_model)$.fitted and augment(lm(total_fatality_rate ~ 1, data = traffic_df))$.fitte
## t = 5.4098e-14, df = 1199, p-value = 1
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1288443  0.1288443
## sample estimates:
## mean of x mean of y
##  18.91856  18.91856
```

3.0.1 Why fit a linear model as a starting point:

We have 1200 observations, which is a reasonably large number of observations. Granted that these observations aren't completely independent of each other. For a given state the year-over-year may bear some correlation. That said, the time frame covers 25years, and this can give enough "spacing" between data

points that there are uncorrelated observations. The data across states can be considered independent. States have high degree of autonomy to set their own rules. Usually, it is the case that a linear model does as good a job as others if we have sufficient amount of data.

3.0.2 Model explanation:

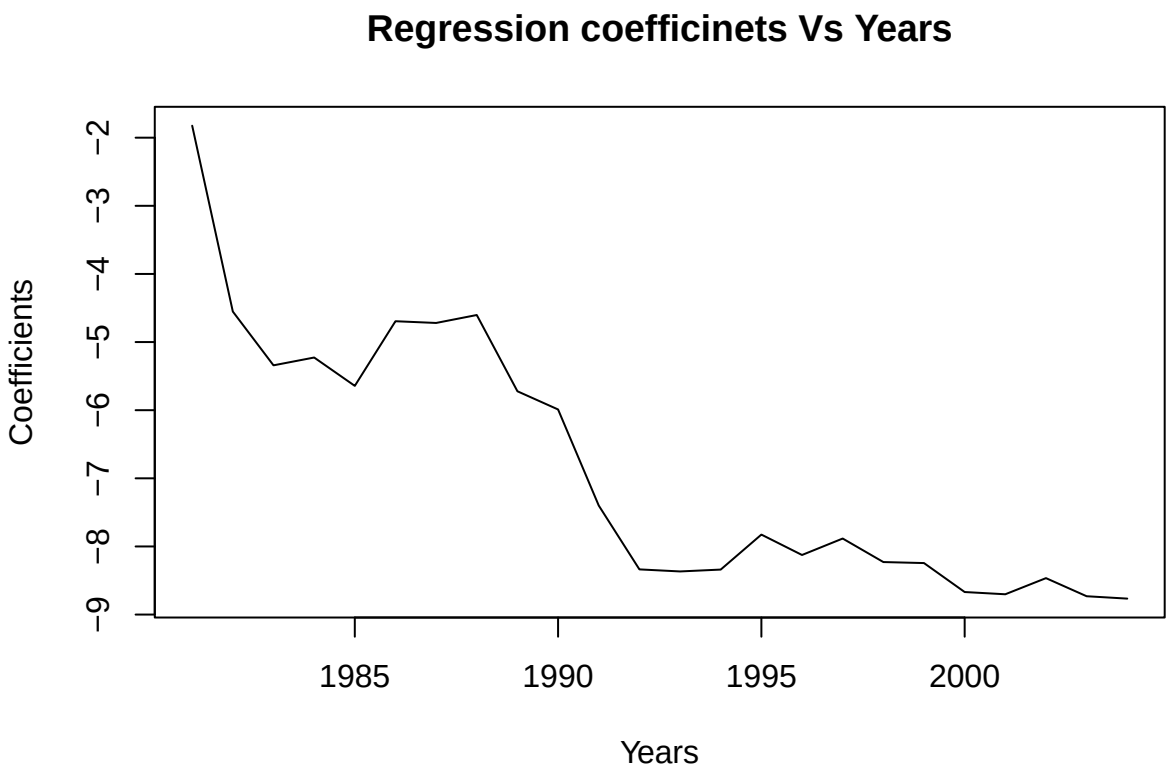
This model basically checks to see if we can fit all the existing data as a linear combination of year dummy variables. In others words, can some linear combination of years able to do a reasonable fit given that such that the sum of difference of squares between predicted and actual values are minimized? A few additional points:

A regression with only Dummy Variables is the ANOVA model. The intercept gives the mean value of traffic fatality rate. The coefficients associated with the dummy variables are not really the "slope" (due to non-continuous variable value); rather it's the differential intercept coefficient, which tells by how the intercept differs going from year to year.

The model output gives most weight to the intercept and it is statistically significant. This shows that irrespective of the year variable there is a base fatality rate.

3.0.3 Did driving become safer over this period?

As we move from 1980 to 2004 we see a reduction in fatality. See the plot below. While the reduction is not monotonic with year the overall trend is toward reduction. Barring year 1981 all other year coefficients are statistically significant. The highest reduction is in 2004 (-8.7656253). We do see an increase from 1985 to 1989. However, for the most part the plot shows that as we move from 1980 toward 2004 we see a reduction in the fatality rate. This could be explained through a combination of technology (better cars, better roads, safety improvements) and, potentially, additional laws. In summary, we see that 2004 has fatality rate reduced by -8.7656253 when compared to the base year of 1980.



3.0.4 Model limitations

The residuals have low mean values ($3.6489265 \times 10^{-15}$). However the Shapiro-Wilk's test doesn't give statistical confidence to assume that the residuals are normally distributed. The R-squared value of the base model is quite low indicating that the model didn't do a good job of capturing all the variance.

The plot of the actual Vs the fitted value shows that we don't have a real good prediction. Above, we also show the result of t-test between the fitted value with dummy variables, and the one without any dummy variable. The result of the t-test shows that there is no difference in the mean.

In mathematical terms, the presence of only dummy regression gives the X matrix (whose columns are the 25 dummy variables with 1200 entries each) full rank (X^TX is invertible). Additionally, the matrix X is orthogonal (i.e the dot product of any two columns is zero). The norm of each column vector is $\sqrt{48} = 4\sqrt{3}$. If we divide matrix X by $4\sqrt{3}$ we get an orthonormal matrix (i.e $X^TX = I$). Thus, the coefficients, $\hat{\beta}$, is given by $\hat{\beta} = (X^TX)^{-1}X^TY$ is indeed X^TY. This means that we can recover the predicted values without any error!! However, this is not what we get in reality. The matrix X is sometimes not even full rank, let alone being orthogonal or orthonormal. Clearly, the model assumption is not rooted in reality.

In summary, the linear model with dummy variables is not a reliable model. Common sense also guides us that the year number alone cannot predict the traffic fatality rate. Given the pure discrete nature of data we can't expect this model to produce unbiased estimate. The same logic extends to the uncertainty estimate. We can't expect this model to recover the residuals correctly. The mathematics behind this, explained above, adds additional reasons why this model is not realistic.

4 (15 points) Expanded Model

Expand the **Preliminary Model** by adding variables related to the following concepts:

- Blood alcohol levels
- Per se laws
- Primary seat belt laws (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.)
- Secondary seat belt laws
- Speed limits faster than 70
- Graduated drivers licenses
- Percent of the population between 14 and 24 years old
- Unemployment rate
- Vehicle miles driven per capita.

If it is appropriate, include transformations of these variables. Please carefully 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 blood alcohol variables defined? Interpret the coefficients that you estimate for this concept.
- Do *per se laws* have a negative effect on the fatality rate?
- Does having a primary seat belt law?

```
expanded_model <- lm(total_fatality_rate ~ d81 + d82 + d83 + d84 + d85 + d86 +
                       d87 + d88 + d89 + d90 + d91 + d92 + d93 + d94 + d95 +
                       d96 + d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04 +
                       blood_alc_lim_08 + blood_alc_lim_10 + adm_lic_revoc_law +
                       primary_seatbelt_law + secondary_seatbelt_law+ sl75 +
                       slnone + grad_drivers_lic_law +
                       percent_pop_aged_14_to_24 + log(unemployment_rate) +
```

```
                      log(miles_driven_per_capita), data = traffic_df)
summary(expanded_model)
```

```
##
## Call:
## lm(formula = total_fatality_rate ~ d81 + d82 + d83 + d84 + d85 +
##     d86 + d87 + d88 + d89 + d90 + d91 + d92 + d93 + d94 + d95 +
##     d96 + d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04 + blood_alc_lim_08 +
##     blood_alc_lim_10 + adm_lic_revoc_law + primary_seatbelt_law +
##     secondary_seatbelt_law + sl75 + slnone + grad_drivers_lic_law +
##     percent_pop_aged_14_to_24 + log(unemployment_rate) + log(miles_driven_per_capita),
##     data = traffic_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.3820  -2.6034  -0.4938   2.3996  20.6481
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -246.75039    7.64654 -32.270  < 2e-16 ***
## d81                            -2.13867    0.80745  -2.649  0.00819 **
## d82                            -6.58330    0.82497  -7.980 3.48e-15 ***
## d83                            -7.57273    0.84218  -8.992  < 2e-16 ***
## d84                            -6.37795    0.85617  -7.449 1.82e-13 ***
## d85                            -7.17077    0.87394  -8.205 6.05e-16 ***
## d86                            -6.59632    0.91076  -7.243 7.98e-13 ***
## d87                            -7.18015    0.94945  -7.562 8.00e-14 ***
## d88                            -7.35654    0.99891  -7.365 3.36e-13 ***
## d89                            -8.94515    1.03770  -8.620  < 2e-16 ***
## d90                           -10.01886    1.06147  -9.439  < 2e-16 ***
## d91                           -12.24746    1.08643 -11.273  < 2e-16 ***
## d92                           -14.13836    1.10653 -12.777  < 2e-16 ***
## d93                           -13.89623    1.12149 -12.391  < 2e-16 ***
## d94                           -13.38752    1.14247 -11.718  < 2e-16 ***
## d95                           -12.85530    1.16818 -11.005  < 2e-16 ***
## d96                           -14.12878    1.19620 -11.811  < 2e-16 ***
## d97                           -13.90690    1.21573 -11.439  < 2e-16 ***
## d98                           -14.40960    1.22581 -11.755  < 2e-16 ***
## d99                           -14.33552    1.24780 -11.489  < 2e-16 ***
## d00                           -14.54898    1.27190 -11.439  < 2e-16 ***
## d01                           -15.60719    1.29776 -12.026  < 2e-16 ***
## d02                           -16.43568    1.31265 -12.521  < 2e-16 ***
## d03                           -16.87547    1.32366 -12.749  < 2e-16 ***
## d04                           -16.39146    1.34780 -12.162  < 2e-16 ***
## blood_alc_lim_08               -2.05594    0.47921  -4.290 1.93e-05 ***
## blood_alc_lim_10               -0.98411    0.35343  -2.784  0.00545 **
## adm_lic_revoc_law              -0.51821    0.29317  -1.768  0.07739 .
## primary_seatbelt_law            0.19417    0.48269   0.402  0.68756
## secondary_seatbelt_law          0.09195    0.41978   0.219  0.82665
## sl75                            3.10700    0.49671   6.255 5.57e-10 ***
## slnone                          6.51571    1.38780   4.695 2.98e-06 ***
## grad_drivers_lic_law           -0.32989    0.52163  -0.632  0.52724
## percent_pop_aged_14_to_24       0.25107    0.11931   2.104  0.03556 *
## log(unemployment_rate)          5.75996    0.47057  12.240  < 2e-16 ***
```

```
## log(miles_driven_per_capita)   28.99403    0.85720  33.824  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.948 on 1164 degrees of freedom
## Multiple R-squared:  0.6267, Adjusted R-squared:  0.6155
## F-statistic: 55.84 on 35 and 1164 DF,  p-value: < 2.2e-16
```

4.0.1 Transformation:

It is not uncommon to see data that are skewed. For instance, certain age group may show higher participation, which then leads to data being skewed to this age group when we look at the data across all age groups. We also see huge variance in data. In these scenarios, log transformation brings a level of normality to the data distribution. As long as the data is positive log transformation is a way of flattening the data, while preserving monotonicity. For instance, x^2 (convex) when log-transformed becomes $2log(x)$ (concave).

The earlier EDA showed the histograms of the data and its log-transformed version. We show below the result of Shapiro-Wilk's test. We chose to transform only the unemployment rate (where log transform makes it normal) and miles driven (where log transform shows considerable improvement). The population percentage doesn't show as much improvement with log transform.

```
var_name <- c()
p_value <- c()

tres <- shapiro.test(log(traffic_df$percent_pop_aged_14_to_24))
var_name <- c(var_name, tres$data.name)
p_value <- c(p_value, tres$p.value)

tres <- shapiro.test(traffic_df$percent_pop_aged_14_to_24)
var_name <- c(var_name, tres$data.name)
p_value <- c(p_value, tres$p.value)

tres <- shapiro.test(log(traffic_df$unemployment_rate))
var_name <- c(var_name, tres$data.name)
p_value <- c(p_value, tres$p.value)

tres <- shapiro.test(traffic_df$unemployment_rate)
var_name <- c(var_name, tres$data.name)
p_value <- c(p_value, tres$p.value)

tres <- shapiro.test(log(traffic_df$miles_driven_per_capita))
var_name <- c(var_name, tres$data.name)
p_value <- c(p_value, tres$p.value)

tres <- shapiro.test(traffic_df$miles_driven_per_capita)
var_name <- c(var_name, tres$data.name)
p_value <- c(p_value, tres$p.value)

data.frame(variable = var_name, p_value)
```

```
##                                   variable      p_value
## 1 log(traffic_df$percent_pop_aged_14_to_24) 4.534606e-13
## 2      traffic_df$percent_pop_aged_14_to_24 3.223496e-17
## 3         log(traffic_df$unemployment_rate) 2.991468e-01
## 4              traffic_df$unemployment_rate 1.265313e-22
```

```
## 5   log(traffic_df$miles_driven_per_capita) 3.426046e-04
## 6        traffic_df$miles_driven_per_capita 9.810094e-15
```

4.0.2 BAC definition and interpretation

The variables "bac08" and "bac10" indicate whether the state had a BAC limit of 0.08% and 0.10% respectively during that year. We transformed "bac08" and "bac10" variables into its binary form represented by "blood_alc_lim_10" and "blood_alc_lim_08". We use the binary form in the regression.

The coefficients of "blood_alc_lim_08" -1.59190 with Std. Error of 0.45346, and that for "blood_alc_lim_10" is -0.62036 with Std. Error of 0.33589. While bac10 reduced the fatality rate it's not significant at 95% confidence level. The bac08 did reduce the fatality and it is statistically significant. It is much further away than two standard deviation. Thus, lower tolerance for blood alcohol content does reduce the law, and the level of 0.08% seems to do a good job of reducing fatalities in a significant way. Common sense thinking supports this result.

4.0.3 Effect of per se law

The per se law codified by the variable "grad_drivers_lic_law" has a coefficient of -0.33880 with Std. Error of 0.52333, and thus not statistically significant. The associated p-value also reveals this. It has a small statistically insignificant negative rate.

4.0.4 Primary seat belt law

This is represented by the variable "primary_seatbelt_law" with coefficient 0.14036 and Std. Error 0.48367 does not have a statistically significant effect (even at 0.1 level). The coefficient is small, indicating that the effect, if at all, is small.

5 (15 points) State-Level Fixed Effects

Re-estimate the **Expanded Model** using fixed effects at the state level.

- What do you estimate for coefficients on the blood alcohol variables? How do the coefficients on the blood alcohol variables change, if at all?
- What do you estimate for coefficients on per se laws? How do the coefficients on per se laws change, if at all?
- What do you estimate for coefficients on primary seat-belt laws? How do the coefficients on primary seatbelt laws change, if at all?

Which set of estimates do you think is more reliable? Why do you think this?

- What assumptions are needed in each of these models?

- Are these assumptions reasonable in the current context?

```
pool.model <- plm(total_fatality_rate ~ year + bac08 + bac10 +
                    adm_lic_revoc_law + primary_seatbelt_law +
                    secondary_seatbelt_law + sl75 + slnone +
                    grad_drivers_lic_law + percent_pop_aged_14_to_24 +
                    log(unemployment_rate) +
                    log(miles_driven_per_capita), data=traffic_df,
                  index=c("state", "year"), model="pooling")
fd.model <- plm(total_fatality_rate ~ year + bac08 + bac10 +
                  adm_lic_revoc_law + primary_seatbelt_law +
                  secondary_seatbelt_law + sl75 + slnone +
                  grad_drivers_lic_law + percent_pop_aged_14_to_24 +
                  log(unemployment_rate) + log(miles_driven_per_capita),
```

```
                data=traffic_df, index=c("state", "year"), model="fd")
between.model <- plm(total_fatality_rate ~ year + bac08 + bac10 +
                       adm_lic_revoc_law + primary_seatbelt_law +
                       secondary_seatbelt_law + sl75 + slnone +
                       grad_drivers_lic_law + percent_pop_aged_14_to_24 +
                       log(unemployment_rate) + log(miles_driven_per_capita),
                     data=traffic_df, index=c("state", "year"),
                     model="between")
within.model <- plm(total_fatality_rate ~ year + bac08 + bac10 +
                      adm_lic_revoc_law + primary_seatbelt_law +
                      secondary_seatbelt_law + sl75 + slnone +
                      grad_drivers_lic_law + percent_pop_aged_14_to_24 +
                      log(unemployment_rate) + log(miles_driven_per_capita),
                    data=traffic_df, index=c("state", "year"),
                    model="within")
```

```
##
## ================================================================================
##                                          Dependent variable:
##                                 ---------------------------------------------
##
##                                   Pooled       FD       Between     Within
##                                     (1)        (2)        (3)        (4)
## --------------------------------------------------------------------------------
## year1981                          -2.140***  -1.225***              -1.571***
##                                    (0.806)    (0.269)               (0.404)
##
## year1982                          -6.591***  -2.918***              -3.471***
##                                    (0.824)    (0.415)               (0.425)
##
## year1983                          -7.550***  -2.858***              -4.092***
##                                    (0.839)    (0.515)               (0.442)
##
## year1984                          -6.286***  -2.322***              -4.668***
##                                    (0.856)    (0.599)               (0.464)
##
## year1985                          -7.050***  -2.175***              -5.184***
##                                    (0.875)    (0.671)               (0.489)
##
## year1986                          -6.460***   -0.642                -4.268***
##                                    (0.913)    (0.733)               (0.527)
##
## year1987                          -7.034***   -0.323                -5.067***
##                                    (0.952)    (0.813)               (0.575)
##
## year1988                          -7.198***    0.160                -5.665***
##                                    (1.001)    (0.899)               (0.632)
##
## year1989                          -8.788***   -0.436                -7.043***
##                                    (1.040)    (0.971)               (0.677)
##
## year1990                          -9.864***   -0.126                -7.153***
##                                    (1.063)    (1.003)               (0.704)
##
```

```
## year1991                    -12.128***   -0.798              -7.856***
##                               (1.087)    (1.015)              (0.719)
##
## year1992                    -13.981***   -1.364              -8.824***
##                               (1.108)    (1.011)              (0.744)
##
## year1993                    -13.764***   -1.231              -9.134***
##                               (1.122)    (0.987)              (0.760)
##
## year1994                    -13.250***   -1.160              -9.572***
##                               (1.143)    (0.955)              (0.782)
##
## year1995                    -12.676***   -0.519              -9.397***
##                               (1.171)    (0.925)              (0.809)
##
## year1996                    -13.960***   -0.461              -9.672***
##                               (1.198)    (0.900)              (0.848)
##
## year1997                    -13.779***   -0.225              -9.843***
##                               (1.216)    (0.844)              (0.873)
##
## year1998                    -14.215***   -0.607              -10.602***
##                               (1.229)    (0.771)              (0.890)
##
## year1999                    -14.128***   -0.587              -10.810***
##                               (1.251)    (0.687)              (0.907)
##
## year2000                    -14.340***   -0.896              -11.387***
##                               (1.275)    (0.620)              (0.923)
##
## year2001                    -15.456***   -0.461              -10.848***
##                               (1.298)    (0.515)              (0.934)
##
## year2002                    -16.328***    0.175              -10.058***
##                               (1.311)    (0.407)              (0.942)
##
## year2003                    -16.746***    0.127              -10.095***
##                               (1.323)    (0.283)              (0.951)
##
## year2004                    -16.188***                       -10.584***
##                               (1.352)                          (0.977)
##
## bac08                        -2.431***   -0.826    -2.103    -1.222***
##                               (0.524)    (0.587)   (2.492)    (0.386)
##
## bac10                        -1.257***  -1.031**   -0.048    -0.912***
##                               (0.387)    (0.445)   (2.083)    (0.263)
##
## adm_lic_revoc_law             -0.461     -0.613     0.346    -1.107***
##                               (0.295)    (0.389)   (1.386)    (0.228)
##
## primary_seatbelt_law           0.192     -0.355    -1.740    -1.159***
##                               (0.482)    (0.482)   (2.726)    (0.337)
##
```

```
## secondary_seatbelt_law          0.079     -0.310     -2.855      -0.228
##                                (0.419)    (0.296)    (2.885)    (0.247)
##
## sl75                           3.080***    -0.366    13.030***  -0.894***
##                                (0.496)    (0.758)    (2.966)    (0.320)
##
## slnone                         6.561***    -2.188     15.324*     0.346
##                                (1.387)    (2.027)    (8.304)    (0.875)
##
## grad_drivers_lic_law           -0.334     -0.184     -0.624      -0.470
##                                (0.521)    (0.369)    (4.268)    (0.290)
##
## percent_pop_aged_14_to_24      0.250**   0.950***     0.088     0.283***
##                                (0.119)    (0.312)    (0.539)    (0.095)
##
## log(unemployment_rate)         5.781***   -1.564***  12.960***  -3.321***
##                                (0.470)    (0.485)    (2.169)    (0.387)
##
## log(miles_driven_per_capita)  28.951***    3.042    27.490***  12.321***
##                                (0.856)    (1.956)    (3.500)    (1.156)
##
## Constant                     -246.298*** -0.188**  -254.071***
##                                (7.639)    (0.095)   (31.418)
##
## ------------------------------------------------------------------------
## Observations                    1,200      1,152       48        1,200
## R2                              0.628      0.178      0.822      0.642
## ========================================================================
## Note:                                        *p<0.1; **p<0.05; ***p<0.01
```

5.1 Blood Alcohol variable coefficient exploration

The variables "bac08" and "bac10" explains the effect of the states implementing the blood alcohol level laws. The Pooled model shows that for states that implemented "bac08" saw a reduction of 2.431 in the fatality rate, and it is statistically significant. The impact of "bac10" on fatality rate is relatively lower at 1.257, and is statistically significant. The "within" model reflects similar results with a bac08/10 coefficient that reduces fatality by 1.222 and 0.912, respectively. The coefficients are lower because we subtract the mean values of the variables in the "within" model. As for First Difference (FD) model, the variable is zero if a state had implemented the law in consecutive years. Thus, the FD tends to minimize the effect of these variables. We get similar results for the coefficients. The reason for bac08 getting higher impact compared to bac10 is due to the number of observations relatively more skewed to 1.0. Finally, for the "between" model we take average for a unit across all time periods, thus the effects are pronounced higher. We expect the model to not show significance likely due to the different times when each state implemented the law. As expected, the standard error is higher for "between" model when compared to other models.

5.2 Per se laws variable coefficient exploration

The variable "adm_lic_revoc_law" (alias "perse") explains the effect of the states implementing the perse law - "suspend or revoke the driving privilege of persons who are arrested for driving with a BAC of . 08% or more". The Pooled model shows that for states that implemented "perse" saw a reduction of 0.461 in the fatality rate, and it is not statistically significant. The "FD" model reflects similar results, with a reduction of 0.613 and the "within" model illustrates significant result with a reduction of 1.107 in fatality rate. As for First Difference (FD) model, the variable is zero if a state had implemented the law in consecutive years. Thus, the FD tends to minimize the effect of these variables. This leads to a coefficient value that

is not statistically significant. Finally, for the "between" model we take average for a unit across all time periods, thus the coefficient shows a positive number with no significance. We expect the model to not show significance likely due to the different times when each state implemented the law. The averaging effect diminished the values of this variable. As expected, the standard error is higher for "between" model when compared to other models.

5.3 Primary seat-belt variable coefficient exploration

The variable "primary_seatbelt_law" explains the effect of the states mandating "driver to wear the seat belt" as a law. The Pooled model shows that for states that implemented "primary_seatbelt_law" saw an increase in fatality of 0.192, but it is not statistically significant. In order for pooled OLS to produce a consistent estimator of β_1, we would have to assume that the unobserved effect, a_i, is uncorrelated with x_{it}. Assuming that correlation between a_i and x_{it} exists we end up with heterogeneity bias or bias caused from omitting a time-constant variable. The "FD" model reflects a reduction of 0.355 and the "within" model illustrates significant result with a reduction of 1.159 in fatality rate. As for First Difference (FD) model, the variable is zero if a state had implemented the law in consecutive years. Thus, the FD tends to minimize the effect of these variables. This leads to a coefficient value that is not statistically significant. Finally, for the "between" model we take average for a unit across all time periods thus the coefficient shows larger number with no significance. As expected the standard error is very high for "between" when compared to other models.

5.4 Reliable model and the reason behind it

First lets explore the assumptions

A1- Linearity: the model is linear in parameters

A2- i.i.d. : The observations are independent across individuals but not necessarily across time. This is guaranteed by random sampling of individuals.

A3- Identifiability: The regressors, including a constant, are not perfectly collinear, and all regressors (but the constant) have non-zero variance and not too many extreme values.

A4- x_it is uncorrelated with idiosyncratic error term u_{it} and individual-specific effect $\gamma_i + E(u_{it}x_{it}) = 0 + E(x_{it}, \gamma_i) = 0$

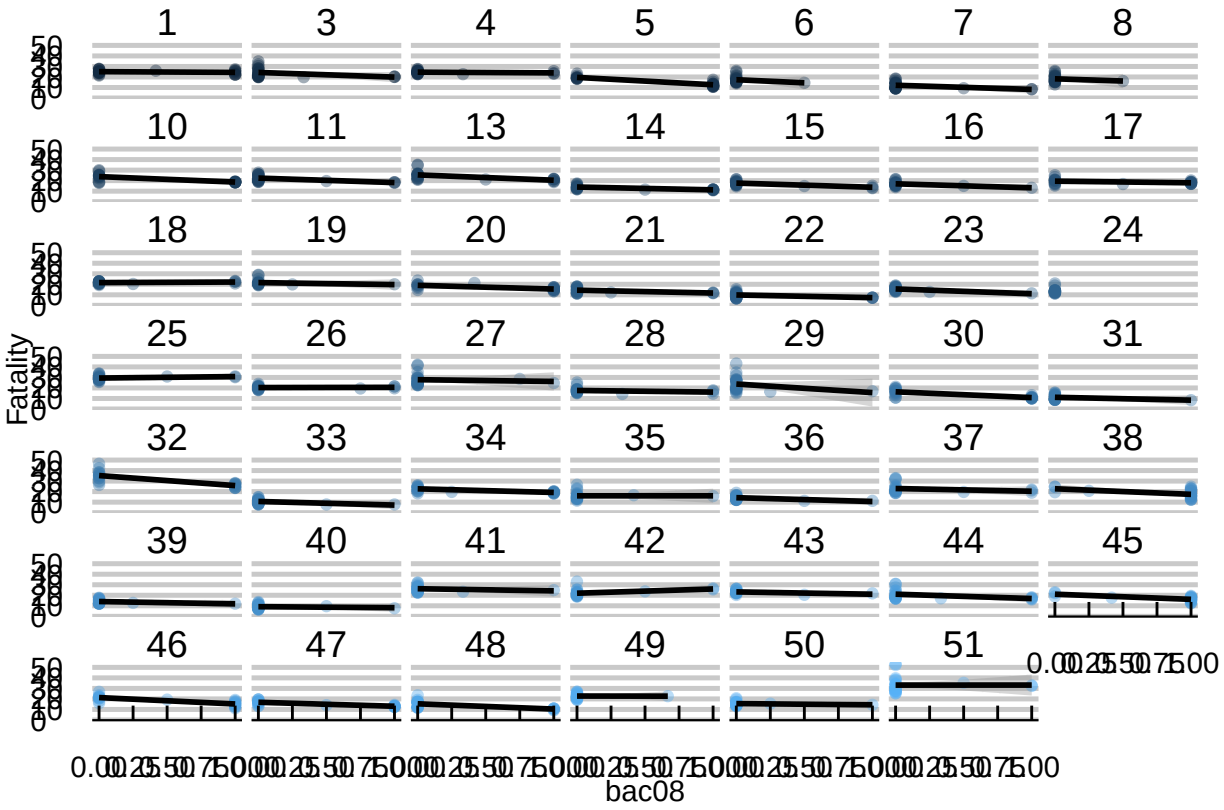
A5- Zero conditional (strict exogeneity) - The most important of these is that δu_i is uncorrelated with δx_i. This assumption holds if the idiosyncratic error at each time t, u_{it}, is uncorrelated with the explanatory variable in both time periods.

For each models the assumptions that should hold true are

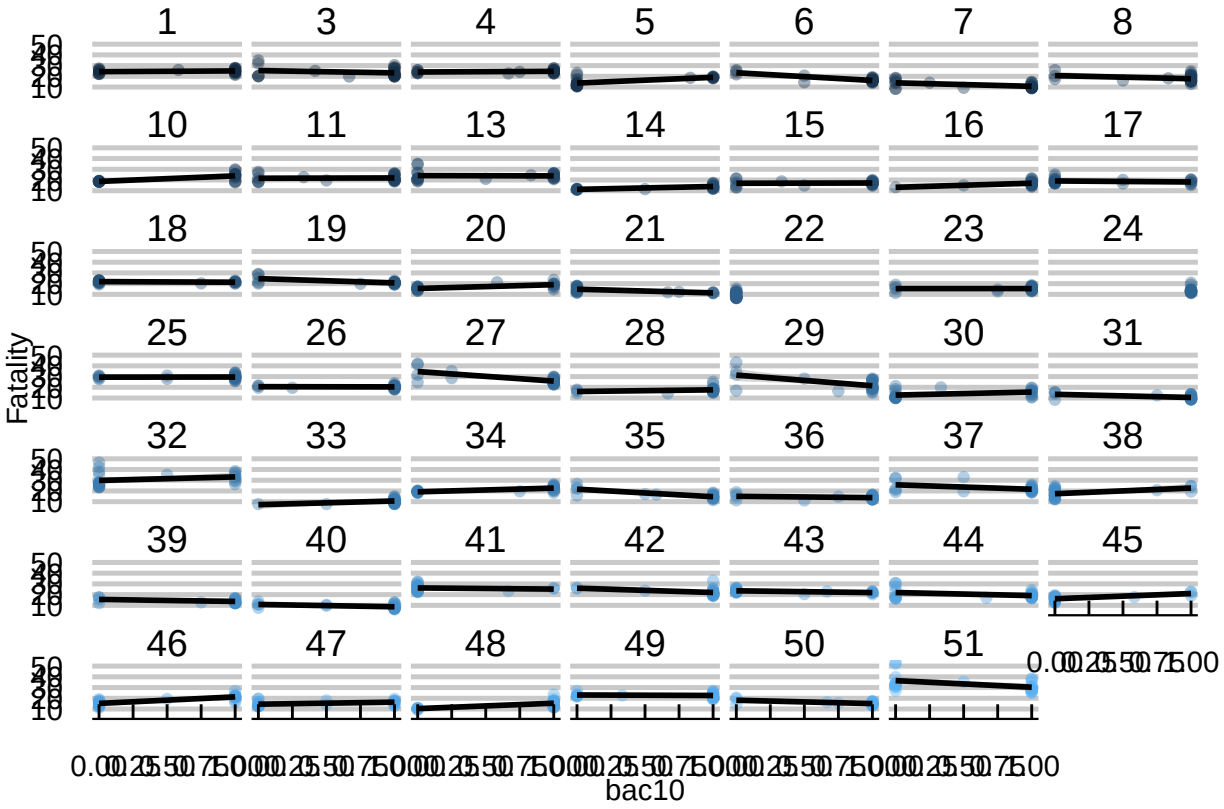
Pooled OLS Model Assumptions - A1,A2,A3,A4 **First Difference Estimator Assumptions** - A1,A2,A3,A5 **Within Model Assumptions** - A1,A2,A3,A5 **Between Model Assumptions** - A1,A2,A3,A5

A1 - Linearity:

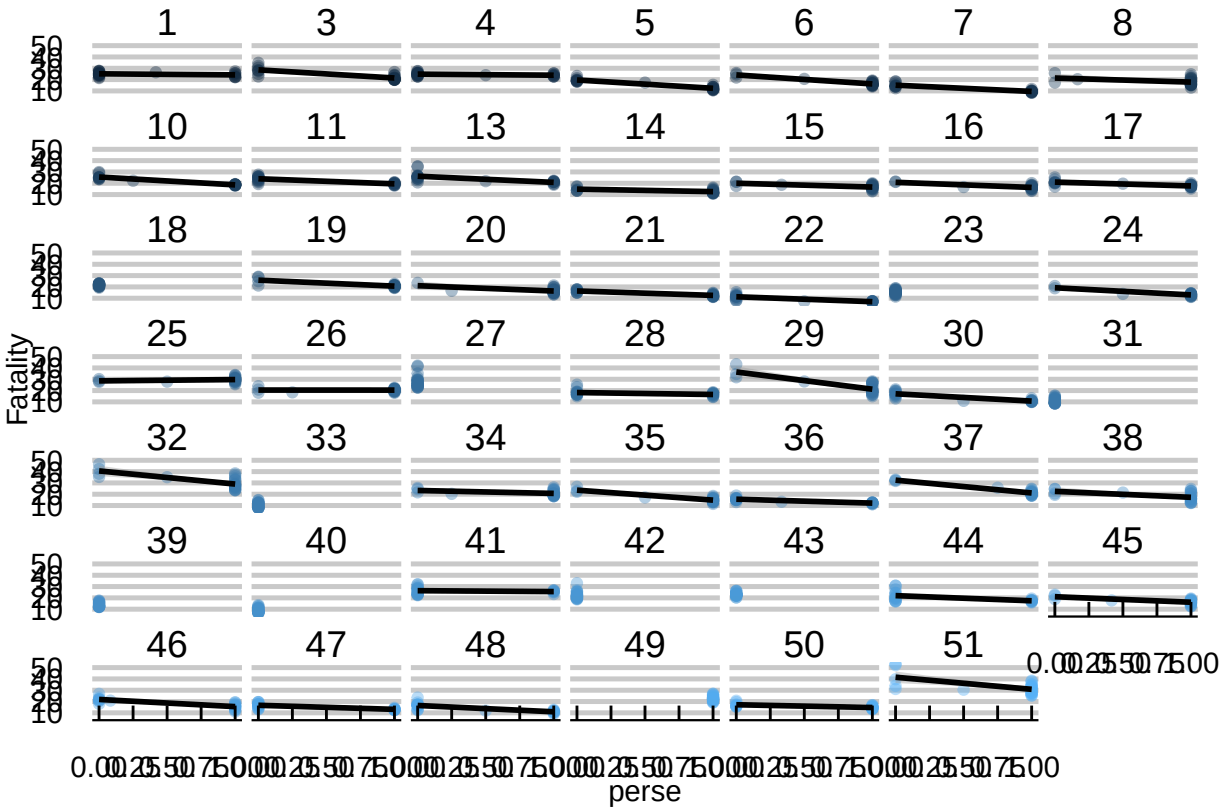
```
## `geom_smooth()` using formula 'y ~ x'
```



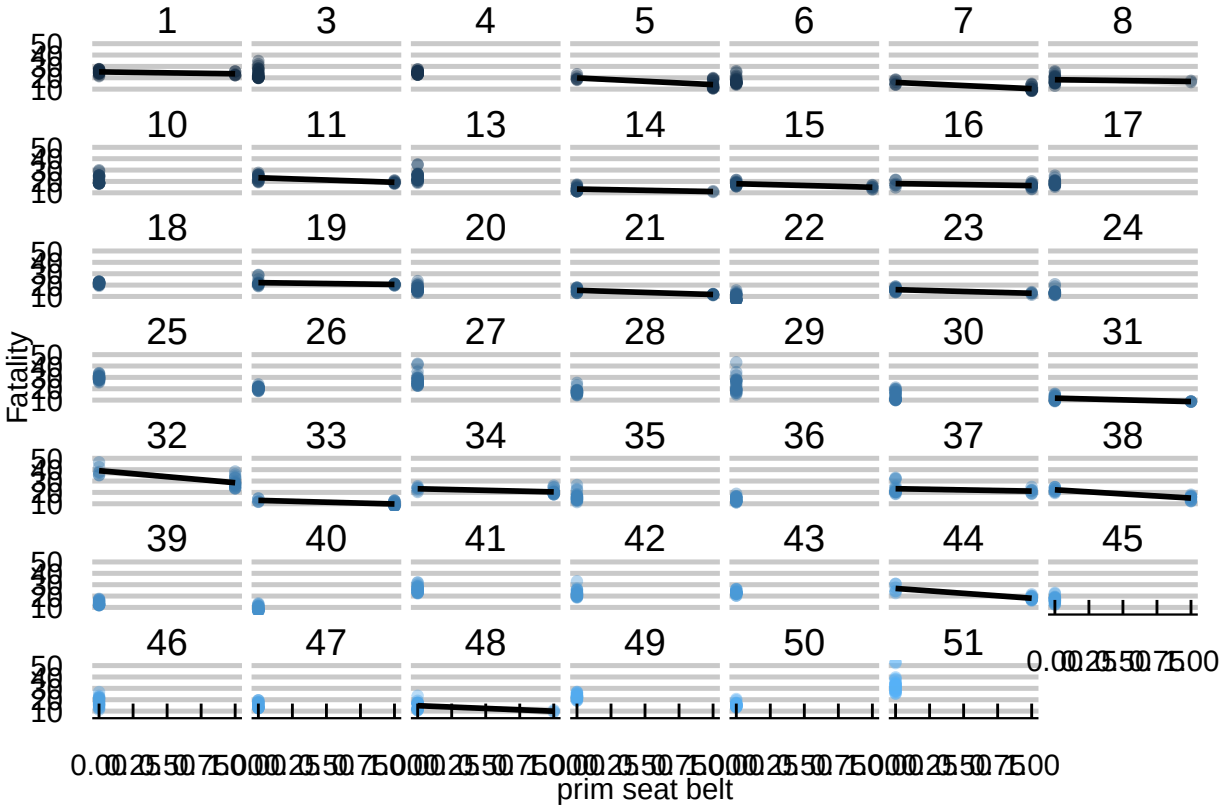

```
## `geom_smooth()` using formula 'y ~ x'
```



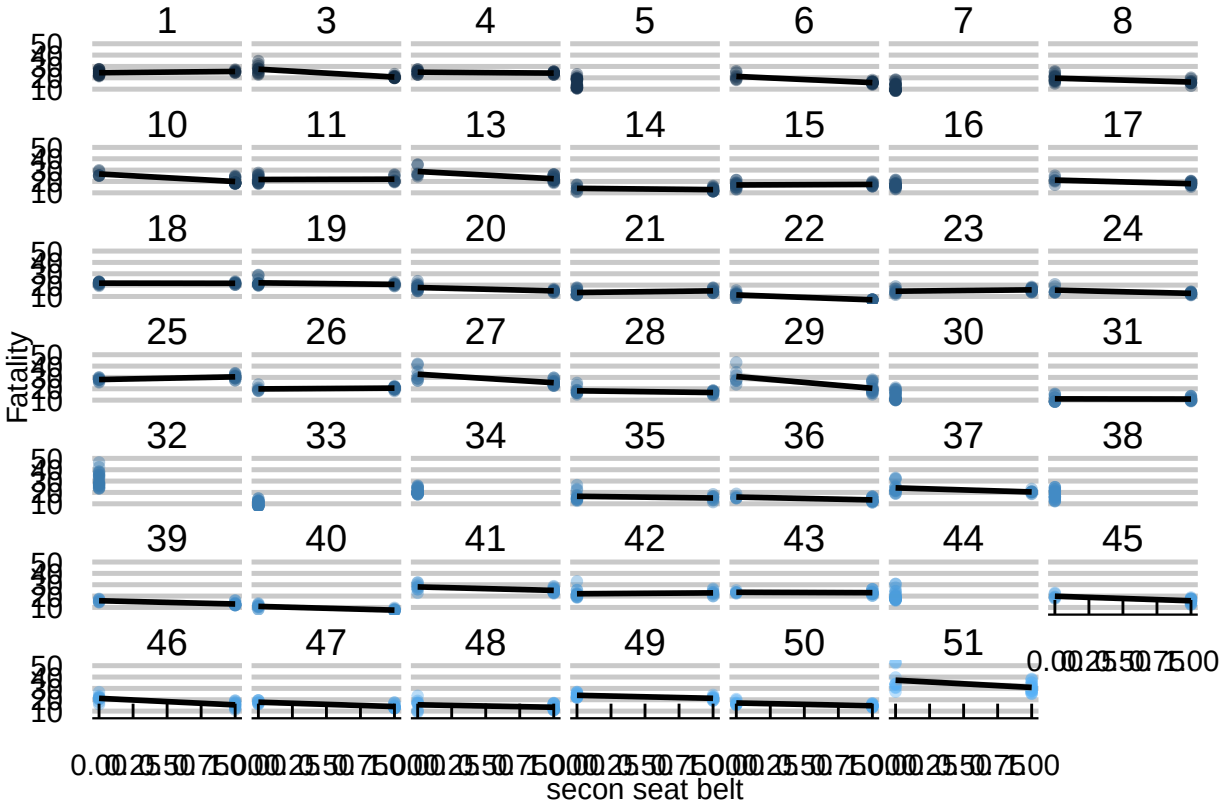

```
## `geom_smooth()` using formula 'y ~ x'
```



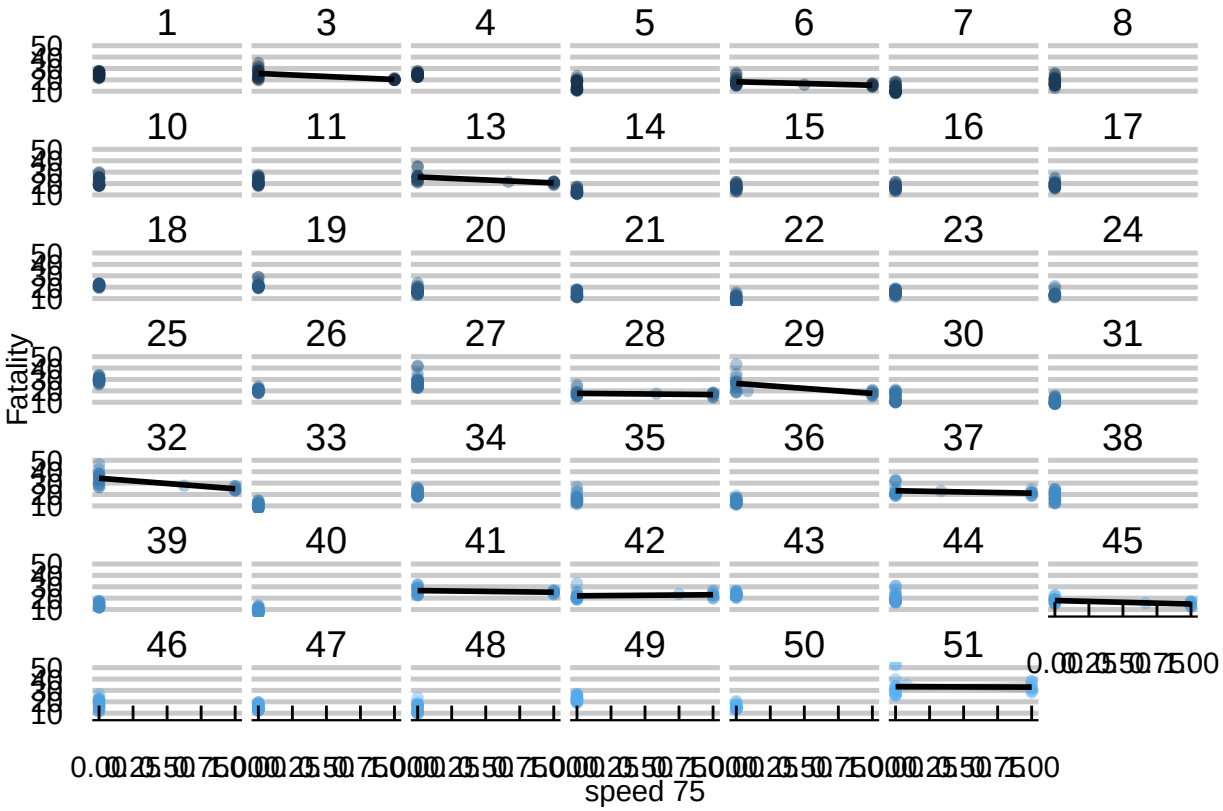

```
## `geom_smooth()` using formula 'y ~ x'
```



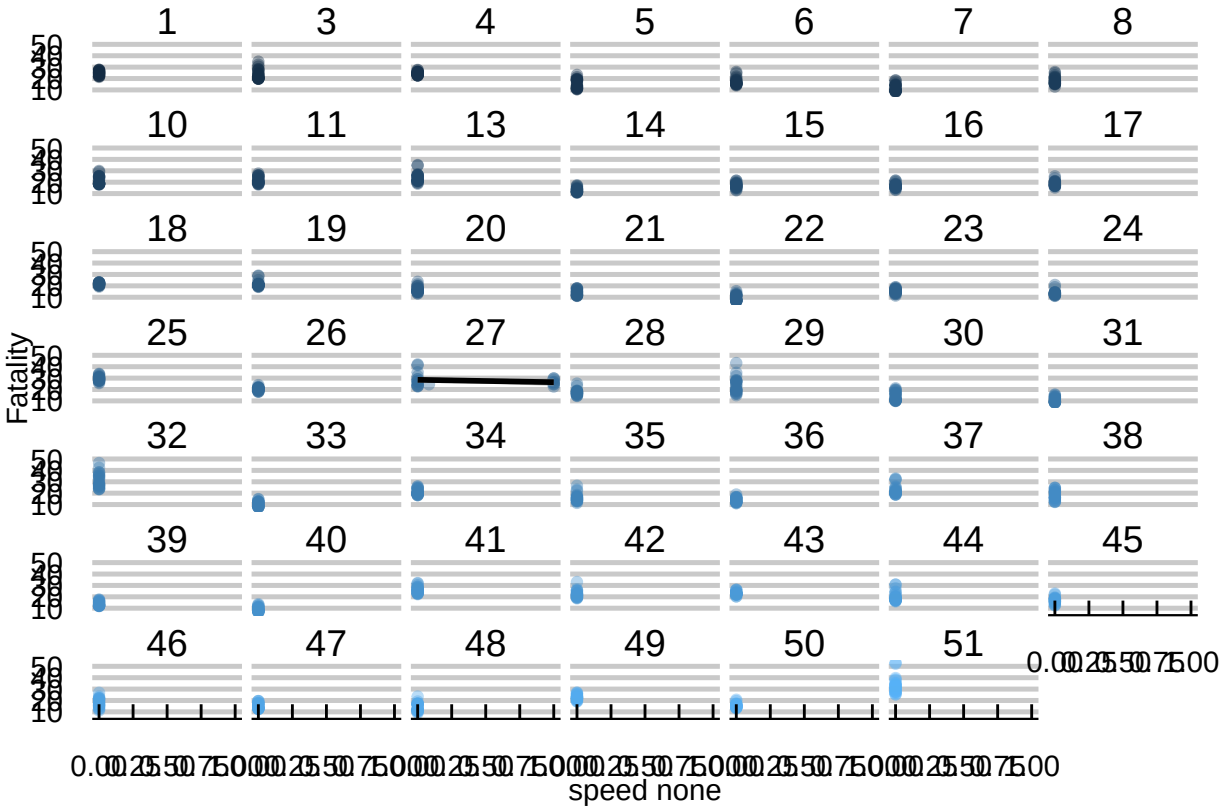
```
## `geom_smooth()` using formula 'y ~ x'
```



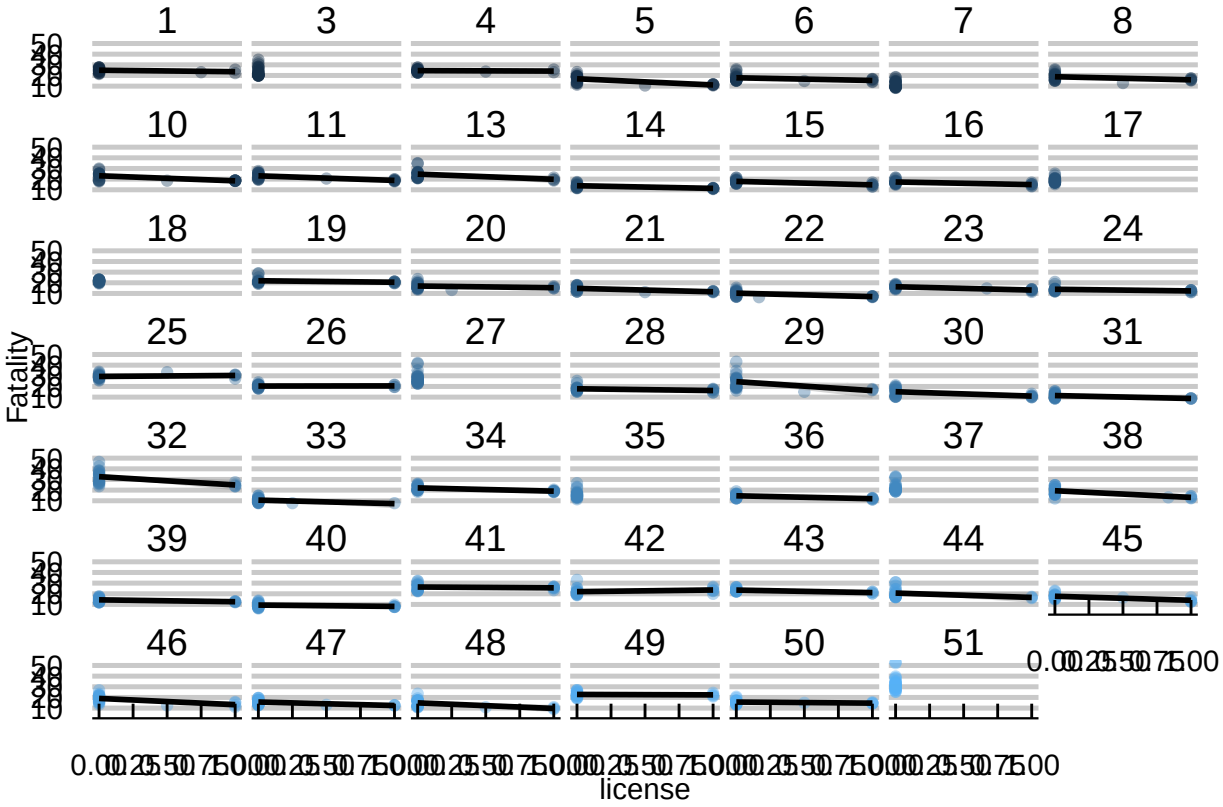

```
## `geom_smooth()` using formula 'y ~ x'
```



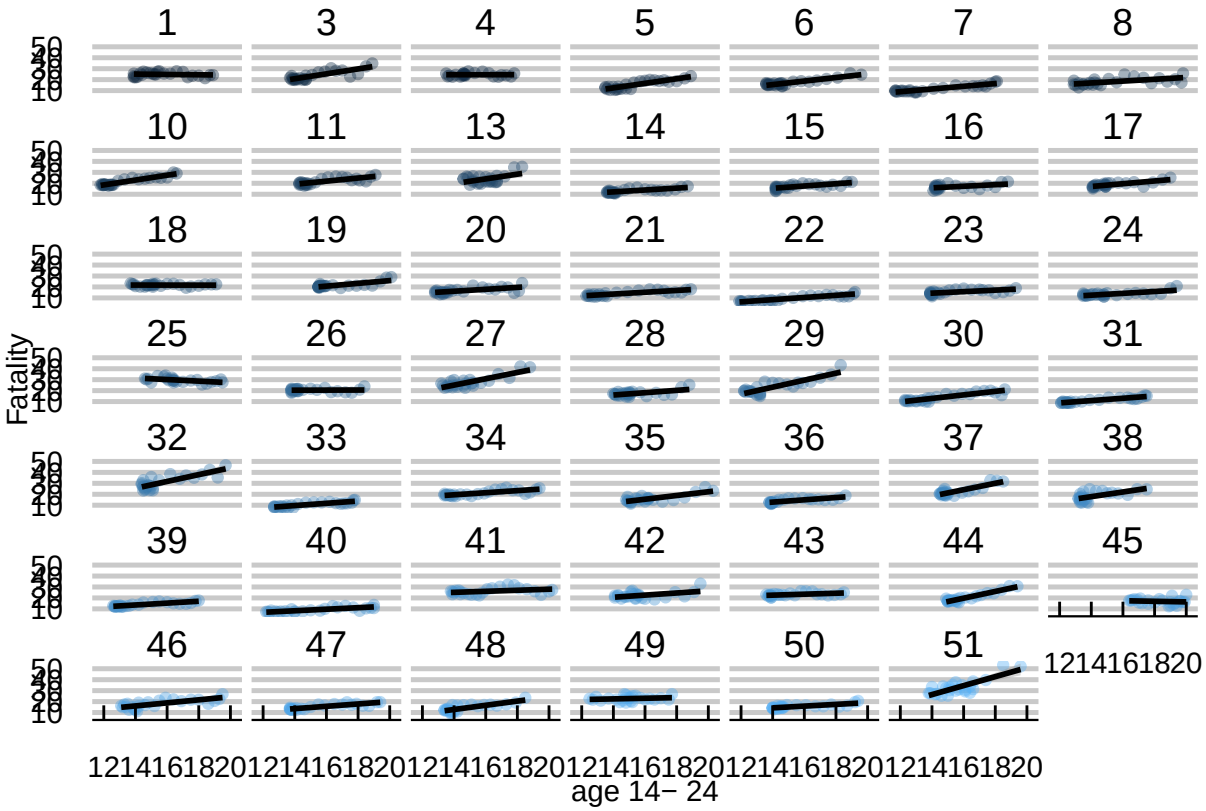

```
## `geom_smooth()` using formula 'y ~ x'
```



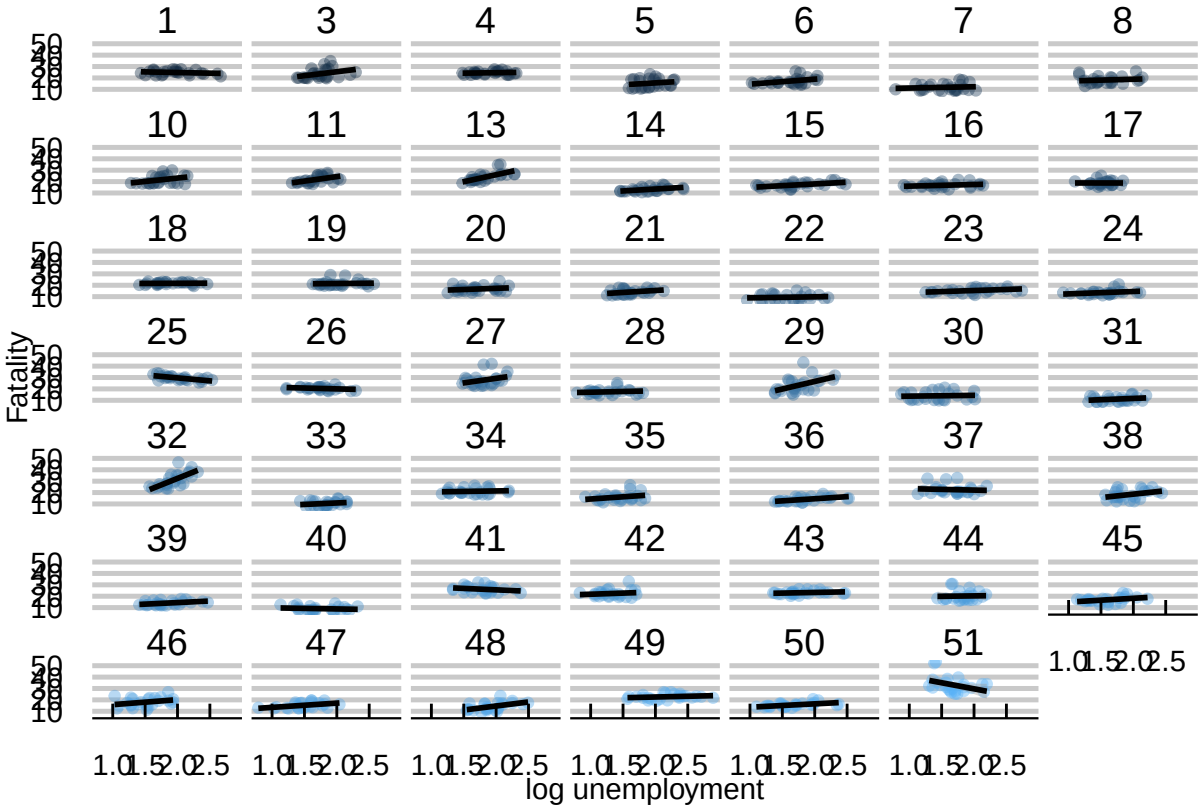
```
## `geom_smooth()` using formula 'y ~ x'
```



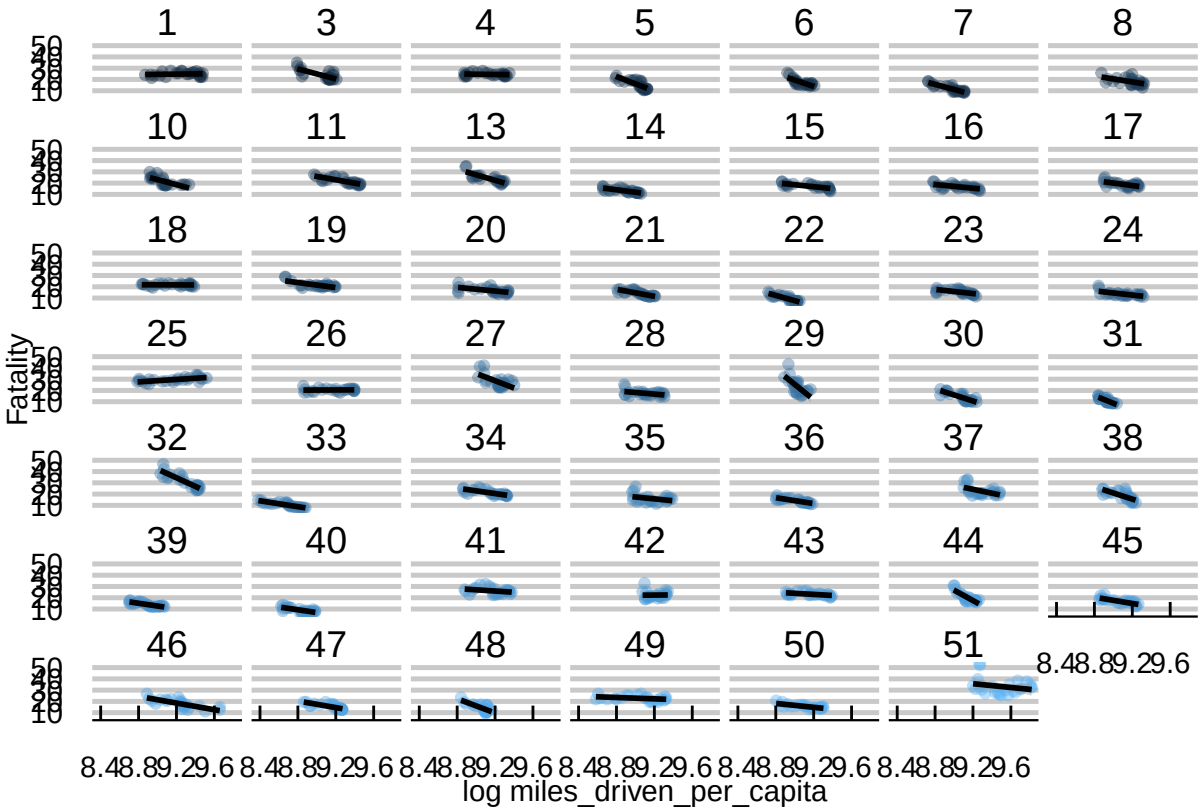

```
## `geom_smooth()` using formula 'y ~ x'
```



```
## `geom_smooth()` using formula 'y ~ x'
```




```
## `geom_smooth()` using formula 'y ~ x'
```



The above plot for each independent **variable** used in the model illustrates the relationship between dependent and the independent variables is linear.

A2 - i.i.d: In the case of all models, except the "between" model we have 1200 data points where a single unit (state in this case) is measured multiple times (across 25 years). Clearly, measurement of a unit repeatedly is not independent. Thus, the 1200 observations are not IID. That said, with 1200 data points we could argue that there are enough observations that are independent. Observations across states qualify for being independent. Observations within an unit across 25 years may also be independent across some of the years.

A3- Identifiability: Proof: The regressors are not perfectly collinear, including a constant

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. A high VIF indicates that the associated independent variable is highly collinear with the other variables in the model. The below result illustrates low VIF (all below 8) the regressors are not perfectly collinear. When we see multiple variables with high VIF score (conventionally, above 8) then the test recommends that we retain one, and drop the others. The VIF scores tell us that these variables are linearly related, and hence can be dropped if we retain only one.

Multicollinearity should not be confused with a raw strong correlation between predictors. What matters is the association between one or more predictor variables, conditional on the other variables in the model. In a nutshell, multicollinearity means that once you know the effect of one predictor (bac08), the value of knowing the other predictor (bac10) is rather low. Thus, one of the predictors doesn't help much in terms of better understanding the model or predicting the outcome given that the other one is part of the model.

```
check_collinearity(pool.model)
```

```
## # Check for Multicollinearity
##
## Low Correlation
##
##                          Term  VIF Increased SE Tolerance
```

```
##                          bac08 3.40         1.84      0.29
##                          bac10 2.57         1.60      0.39
##             adm_lic_revoc_law 1.63         1.28      0.62
##           primary_seatbelt_law 2.64         1.62      0.38
##         secondary_seatbelt_law 3.38         1.84      0.30
##                           sl75 1.37         1.17      0.73
##                         slnone 1.11         1.05      0.90
##           grad_drivers_lic_law 2.93         1.71      0.34
##      percent_pop_aged_14_to_24 3.86         1.96      0.26
##         log(unemployment_rate) 1.90         1.38      0.53
##  log(miles_driven_per_capita) 2.22         1.49      0.45
##
## High Correlation
##
##  Term   VIF Increased SE Tolerance
##  year 43.23         6.58      0.02
```

Proof: The regressors have non-zero variance and not too many extreme values (but constant). Based on the below variance information, the independent variables have non-zero variance and not too many extreme values.

```
var(traffic_df$bac08)
```

```
## [1] 0.1604773
```

```
var(traffic_df$bac10)
```

```
## [1] 0.2229718
```

```
var(traffic_df$adm_lic_revoc_law)
```

```
## [1] 0.2429163
```

```
var(traffic_df$sl75)
```

```
## [1] 0.07241937
```

```
var(traffic_df$slnone)
```

```
## [1] 0.007454661
```

```
var(traffic_df$grad_drivers_lic_law)
```

```
## [1] 0.1401044
```

```
var(traffic_df$percent_pop_aged_14_to_24)
```

```
## [1] 3.523916
```

```
var(log(traffic_df$unemployment_rate))
```

```
## [1] 0.1115752
```

```
var(log(traffic_df$miles_driven_per_capita))
```

```
## [1] 0.0392362
```

A4- $x_i t$ is uncorrelated with idiosyncratic error term u_{it} and individual-specific effect γ_i

Durbin-Watson and Breusch-Godfrey/Wooldridge test (with order 2) are used to validate this.

```
pdwtest(pool.model)
```

```
##
##  Durbin-Watson test for serial correlation in panel models
##
## data:  total_fatality_rate ~ year + bac08 + bac10 + adm_lic_revoc_law +  ...
## DW = 0.42628, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```
pbgtest(pool.model, order = 2)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel models
##
## data:  total_fatality_rate ~ year + bac08 + bac10 + adm_lic_revoc_law +  ...
## chisq = 752.63, df = 2, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```
pdwtest(fd.model)
```

```
##
##  Durbin-Watson test for serial correlation in panel models
##
## data:  total_fatality_rate ~ year + bac08 + bac10 + adm_lic_revoc_law +  ...
## DW = 2.7277, p-value = 0.7605
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```
pbgtest(fd.model, order = 2)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel models
##
## data:  total_fatality_rate ~ year + bac08 + bac10 + adm_lic_revoc_law +  ...
## chisq = 161.9, df = 2, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```
pdwtest(within.model)
```

```
##
##  Durbin-Watson test for serial correlation in panel models
##
## data:  total_fatality_rate ~ year + bac08 + bac10 + adm_lic_revoc_law +  ...
## DW = 1.067, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```
pbgtest(within.model, order = 2)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel models
##
## data:  total_fatality_rate ~ year + bac08 + bac10 + adm_lic_revoc_law +  ...
## chisq = 282.94, df = 2, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```
pdwtest(between.model)
```

```
##
##  Durbin-Watson test for serial correlation in panel models
##
## data:  total_fatality_rate ~ year + bac08 + bac10 + adm_lic_revoc_law +  ...
```

```
## DW = 1.6435, p-value = 0.09671
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```
pbgtest(between.model, order = 2)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel models
##
## data:  total_fatality_rate ~ year + bac08 + bac10 + adm_lic_revoc_law +  ...
## chisq = 5.1005, df = 2, p-value = 0.07806
## alternative hypothesis: serial correlation in idiosyncratic errors
```

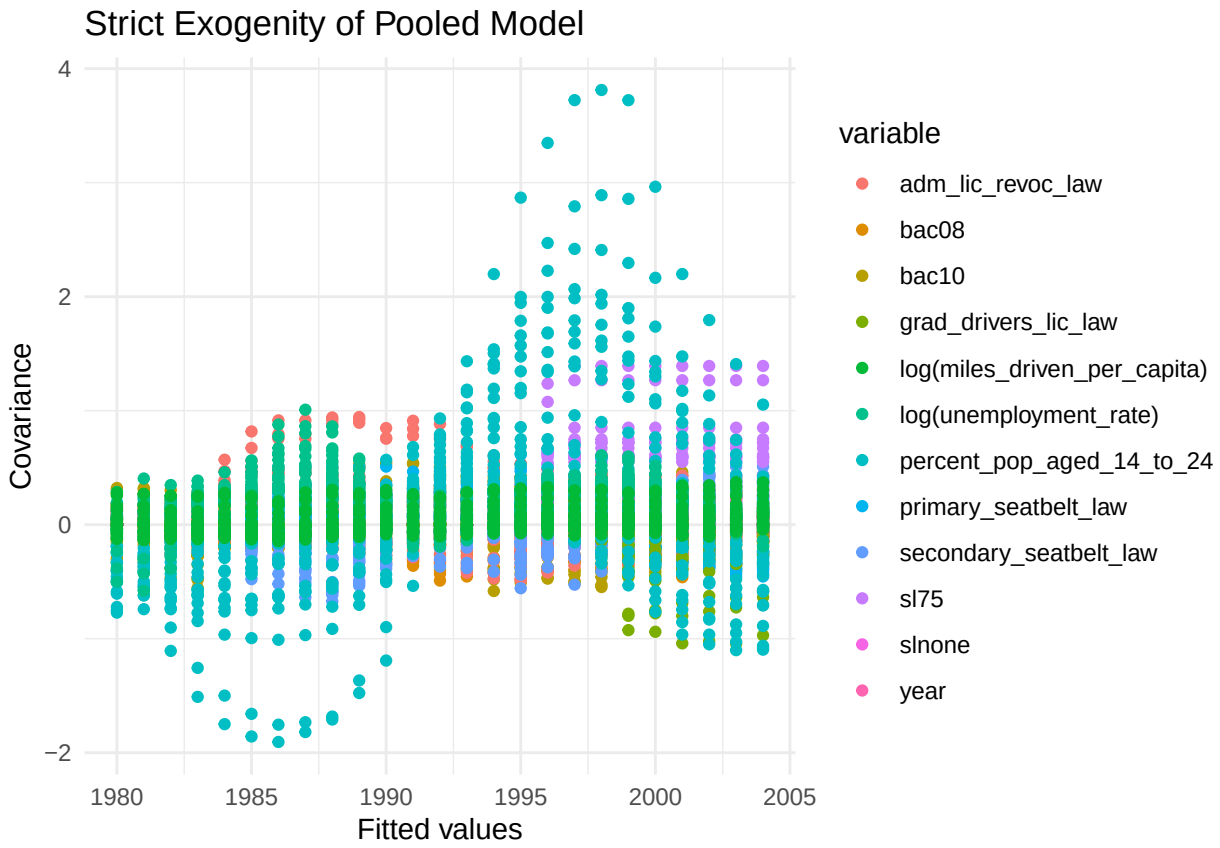
We don't have evidence to reject the null hypothesis of no serial correlation in the between models. We do have evidence to reject the null hypothesis of no serial correlation in the within and pooled OLS models. For first difference estimator model, Durbin-Watson and Breusch-Godfrey/Wooldridge test shows contradictory results. The first difference model usually ends up with correlated error terms, except when the error is random walk. For a generic AR(1) process we do expect to see serial correlation.

A5- Zero conditional (strict exogeneity) - The most important of these is that δu_i is uncorrelated with δx_i.

```
df <- traffic_df
df$`log(unemployment_rate)` <- log(df$unemployment_rate)
df$`log(miles_driven_per_capita)` <- log(df$miles_driven_per_capita)
df$fit_model <- augment(pool.model)$.fitted
df$resid_model <- residuals(pool.model)

df_model <- data.frame(matrix(ncol = 3, nrow = 0))
x <- c("year", "variable", "covariance")
colnames(df_model) <- x

for (t in unique(df$year)){
  for (var_name in colnames(pool.model$model)[c(-1)]){
    for (s in unique(df$year)){
      c = cov(df[df$year == t,][var_name], df[df$year == s,]$resid_model)
      df_model <- rbind(df_model,data.frame(year = t, variable = var_name,
                                            covariance = c[1]))
    }
  }
}
```



Significant number of correlations are non-zero, which means that the assumption of lack of correlation is incorrect. Hence the estimates are not unbiased; the model may be consistent, though. EDA showed that there was an overall increasing trend in miles driven and an overall decreasing trend in total fatality rate. The plot above shows that there is a violation in the strict exogeneity assumption. However we chose to include the per-capita miles driven variable as miles driven has to be an explanatory variable in predicting total fatality rate (if miles driven is zero the total fatalities from auto accidents would be zero too).

5.5 Verdict

Pooled model uses all the observations as if these are independently measured. In general, for panel data, we know that this assumption is not true. If assumptions $A1 - A4$ are valid then pooled OLS will give consistent results. For the data set we have these assumptions don't hold good - in particular the IID assumption (A2) is not valid. Thus, pooled estimates are not unbiased and likely not consistent either. We must recognize that the pooled model had $R^2 = 0.628$, which means that the explanatory variables do capture most of the variance. This shouldn't be a surprise given that we have 1200 observations. As the number of observations go up the pooled OLS comes close to "within" or "between" models.

The Fixed Effect (FE) model with first difference gets rid of the fixed effects (and its correlation with the explanatory variables). However, the error difference ($\Delta u_i = u_{i2} - u_{i1}$) may not be uncorrelated with all explanatory variables. The error term u_{i1} is guaranteed to be uncorrelated to all x_{i1} but may be correlated to x_{i2}. In our case, unobserved factors such as increased highway patrol may remain the same or vary across time. Hence, the first difference model, again, is not likely to give unbiased or consistent estimate. The model has $R^2 = 0.178$, which means that the model has left a lot of variance in the residuals, instead of accounting for in the explanatory variables.

The "within" model has $R^2 = 0.642$. This model, which includes the de-meaned values of the outcome and the explanatory variables, seems to perform well. The result of de-mean action removed the fixed effect

variables and preserves the error terms uncorrelated characteristic with the explanatory variables. The residuals seem to approximate a normal distribution, shown in the QQ-Plot below. This model is probably a tad better than the pooled OLS but it is very close.

The "between" model has $R^2 = 0.822$ and does the best job of capturing most of the variance in the model. The QQ-Plot also shows adherence to a normal distribution. The number of observations are only 48, as expected. This model does a good job because the average of each state across 25 years are likely independent of each other. In general, observation across each state tend to be independent. By and large each state is autonomous in setting driving regulations.

Based on the pFtest result, p-value of 0.002243, we reject the null hypothesis of no fixed effects. This means we should include state and/or time fixed effects in our model. The QQ-Plot for the "between" model gives the best results along with the R^2 values. Thus, we choose the "between" model as the most optimal one.

Comparing both the Wooldridge's hypothesis tests, we see that the both the models (FD and FE) suffer from serial correlation. This suggests that both the models are not good and we need to fit a better model.

```
pwfdtest(fd.model, data=traffic_df, index=c("state","year"), h0="fe")
```

```
##
##  Wooldridge's first-difference test for serial correlation in panels
##
## data:  fd.model
## F = 8.9208, df1 = 1, df2 = 1102, p-value = 0.002882
## alternative hypothesis: serial correlation in original errors
```

```
pwfdtest(fd.model, data=traffic_df, index=c("state","year"), h0="fd")
```

```
##
##  Wooldridge's first-difference test for serial correlation in panels
##
## data:  fd.model
## F = 94.307, df1 = 1, df2 = 1102, p-value < 2.2e-16
## alternative hypothesis: serial correlation in differenced errors
```

```
pFtest(between.model,pool.model)
```

```
##
##  F test for individual effects
##
## data:  total_fatality_rate ~ year + bac08 + bac10 + adm_lic_revoc_law +  ...
## F = 2.1951, df1 = 1128, df2 = 36, p-value = 0.002243
## alternative hypothesis: significant effects
```

```
pFtest(within.model, pool.model)
```

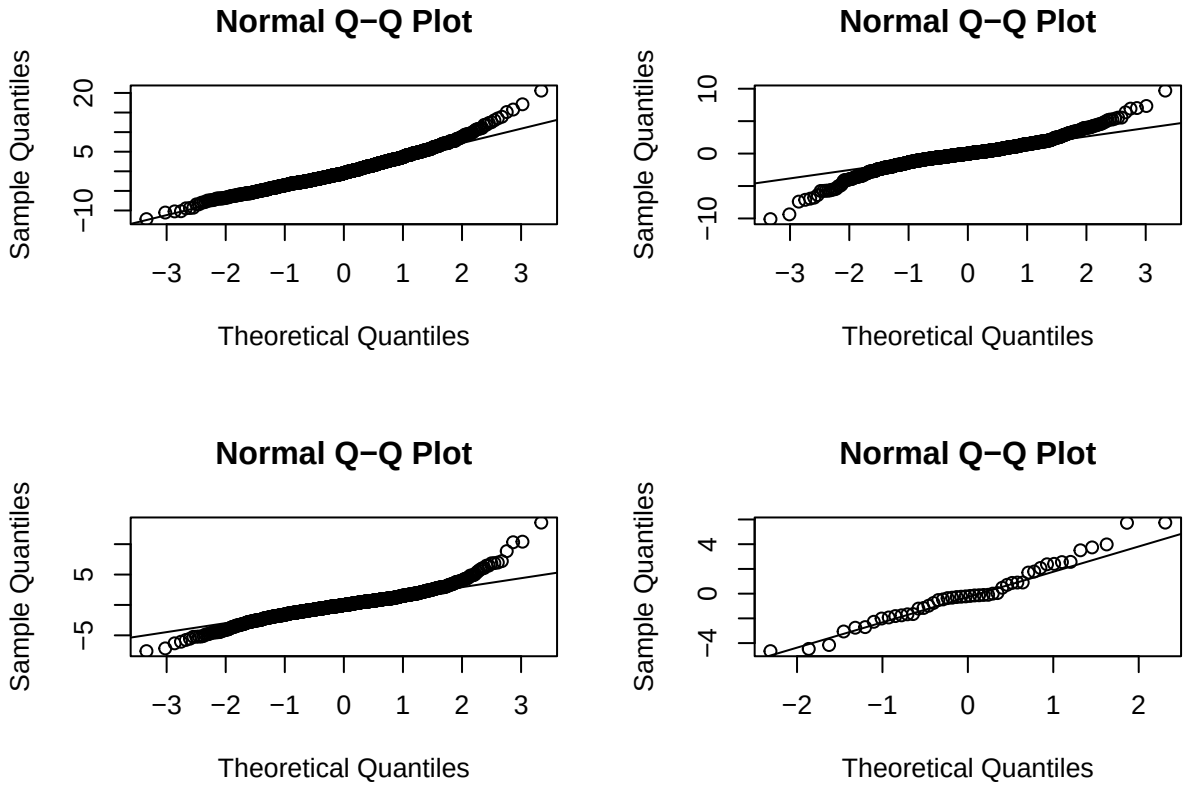
```
##
##  F test for individual effects
##
## data:  total_fatality_rate ~ year + bac08 + bac10 + adm_lic_revoc_law +  ...
## F = 75.242, df1 = 47, df2 = 1117, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

```
par(mfrow=c(2,2))
qqnorm(pool.model$residuals)
qqline(pool.model$residuals)

qqnorm(fd.model$residuals)
qqline(fd.model$residuals)
```

```
qqnorm(within.model$residuals)
qqline(within.model$residuals)

qqnorm(between.model$residuals)
qqline(between.model$residuals)
```



6 (10 points) Consider a Random Effects Model

Instead of estimating a fixed effects model, should you have estimated a random effects model?

- Please state the assumptions of a random effects model, and evaluate whether these assumptions are met in the data.
- If the assumptions are, in fact, met in the data, then estimate a random effects model and interpret the coefficients of this model. Comment on how, if at all, the estimates from this model have changed compared to the fixed effects model.
- If the assumptions are **not** met, then do not estimate the data. But, also comment on what the consequences would be if you were to *inappropriately* estimate a random effects model. Would your coefficient estimates be biased or not? Would your standard error estimates be biased or not? Or, would there be some other problem that might arise?

We will start by setting the model and then testing the assumptions needed for the random effect model.

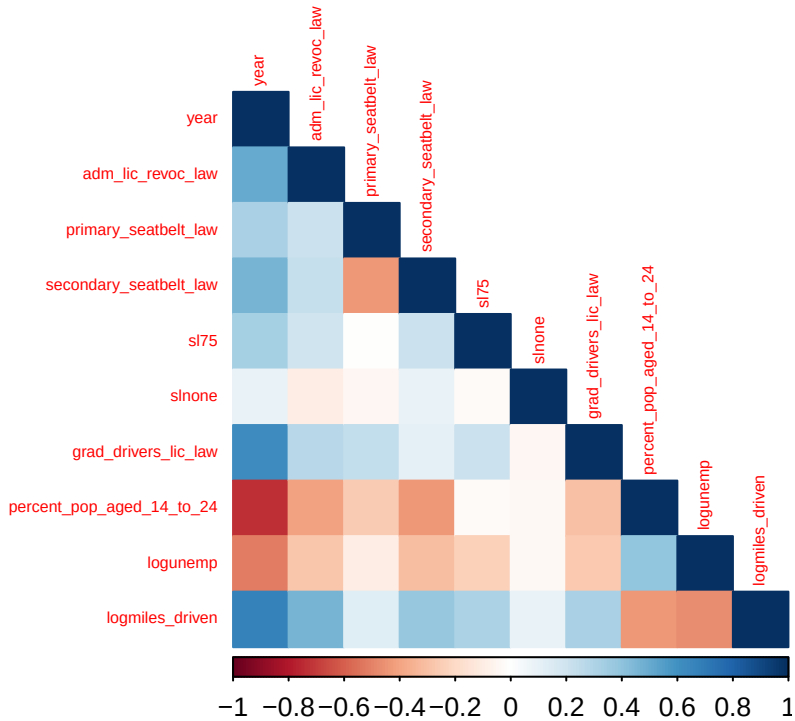
```
re.model <- plm(total_fatality_rate ~ year+bac10 + adm_lic_revoc_law + primary_seatbelt_law +
                  secondary_seatbelt_law + sl75 + slnone +
                  grad_drivers_lic_law + percent_pop_aged_14_to_24 +
                  log(unemployment_rate) + log(miles_driven_per_capita),
                data=traffic_df, index=c("state", "year"), model="random")
```

The first assumption of the random effects model is that there are no perfect linear relationships among the

explanatory variables.

```
colin_data <- traffic_df
colin_data$logmiles_driven <- log(colin_data$miles_driven_per_capita)
colin_data$logunemp <- log(colin_data$unemployment_rate)

colin_data = colin_data %>% select(year,adm_lic_revoc_law, primary_seatbelt_law,secondary_seatbelt_law,s
                                   percent_pop_aged_14_to_24,logunemp,
                                   logmiles_driven)
corrplot(cor(colin_data), method="color", number.cex=0.5, tl.cex=0.5,
         type='lower')
```



From the above correlation plot we see some concerning correlations, particularly between percent of population aged 14 to 24 and the year variable as well as the log of unemployment and the year. These high correlations may indicate multicollinearity (or if there is a perfect linear relationship between variables).

Additionally, we will examine the Variance Inflation Factor matrix of the model for multicollinearity.

```
car::vif(re.model)
```

```
##                                    GVIF Df GVIF^(1/(2*Df))
## year                         173.149222 24        1.113355
## bac10                          1.457606  1        1.207314
## adm_lic_revoc_law              2.096728  1        1.448008
## primary_seatbelt_law           2.471865  1        1.572216
## secondary_seatbelt_law         3.382248  1        1.839089
## sl75                           1.577961  1        1.256169
## slnone                         1.116779  1        1.056778
## grad_drivers_lic_law           3.333315  1        1.825737
## percent_pop_aged_14_to_24      7.902704  1        2.811175
## log(unemployment_rate)         3.303485  1        1.817549
## log(miles_driven_per_capita)   7.685368  1        2.772250
```

We see high values for year=(1995, 1996, 2000, 2003, 2004) percent_pop_aged_14_to_24, and log(miles_driven_per_capita) indicating the possible presence of multicollinearity in these variables.

Assessing our analysis of multicollinearity in the model explanatory variables, we have not provided sufficient evidence in supporting the lack of perfect multicollinearity between the variables as there are some strong correlations (linear relationships) between certain variables.

The second assumption is that there is no correlation between the unobserved effect (random and fixed effects) and the explanatory variables. Using a random effects model imposes the error structure that the error term v_{it} is equal to the sum of variation between groups and variation within groups onto the model residuals, allowing to properly specify the residuals and more efficiently estimate the coefficients of interest. This requires the assumption of independence between random effects and the other predictors in the model. The assumptions for the fixed effect model are discussed above, the additional assumption of independence of random effects and other predictors in the model is evaluated below. The test we run is the Hausman Test for fixed versus random effects. The null hypothesis is that the random effects model is acceptable while the alternative hypothesis is that there is correlation between residuals and predictors, meaning that we should use the FE model.

```
phtest(between.model, re.model)
```

```
##
##  Hausman Test
##
## data:  total_fatality_rate ~ year + bac08 + bac10 + adm_lic_revoc_law +  ...
## chisq = 109.22, df = 10, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

```
#traffic.new.row
```

The Hausman test results in a p-value of `phtest(between.model, re.model)$p.value` which is far above the standard 0.05 value used to determine statistical significance. In this case, we do not have support to reject the null hypothesis and are thus can assume that we should accept the null hypothesis that the random effects model is acceptable as both models are consistent, and we know that the random effects model is more efficient, and thus the preferable model to use.

The third assumption is that of homoskedastic errors, which we can test below using the Pesaran's CD test:

```
pcdtest(re.model)
```

```
##
##  Pesaran CD test for cross-sectional dependence in panels
##
## data:  total_fatality_rate ~ year + bac10 + adm_lic_revoc_law + primary_seatbelt_law +     secondary
## z = -0.46055, p-value = 0.6451
## alternative hypothesis: cross-sectional dependence
```

While the p-value is not statistically significant indicating that we do not have support to reject the null hypothesis in favor of the alternative hypothesis of cross-sectional dependence, we have already violated assumption 2 and likely assumption 1, and thus should not proceed with this model.

Overall, the main assumption of the RE model, assumption 2, was violated by showing that the RE model is inconsistent. Additionally, there are concerns of multicollinearity in the data. From this, we will not proceed with estimating the random effects model.

If we were to inappropriately estimate a random effects model, we would be incorrectly assuming that the random effects and other predictors are independent of one another. This would lead to omitted variable bias as the correlation between the random effects and the explanatory variables of interest would not allow for accurate estimation of the coefficient. Standard errors will also be biased as we are assuming that the random effects, which are included in the error term, are incorrectly uncorrelated with the predictors - given that there is correlation, this will introduce bias into the standard errors.

7 (10 points) Model Forecasts

The COVID-19 pandemic dramatically changed patterns of driving. Find data (and include this data in your analysis, here) that includes some measure of vehicle miles driven in the US. Your data should at least cover the period from January 2018 to as current as possible. With this data, produce the following statements:

- Comparing monthly miles driven in 2018 to the same months during the pandemic:
  - What month demonstrated the largest decrease in driving? How much, in percentage terms, lower was this driving?
  - What month demonstrated the largest increase in driving? How much, in percentage terms, higher was this driving?

Now, use these changes in driving to make forecasts from your models.

- Suppose that the number of miles driven per capita, increased by as much as the COVID boom. Using the FE estimates, what would the consequences be on the number of traffic fatalities? Please interpret the estimate.
- Suppose that the number of miles driven per capita, decreased by as much as the COVID bust. Using the FE estimates, what would the consequences be on the number of traffic fatalities? Please interpret the estimate.

```
setwd("/home/rstudio/kumarn/MIDS/w271/Lab-3")
vehicle_miles <- read.csv('./data/vehiclemiles.csv')
```

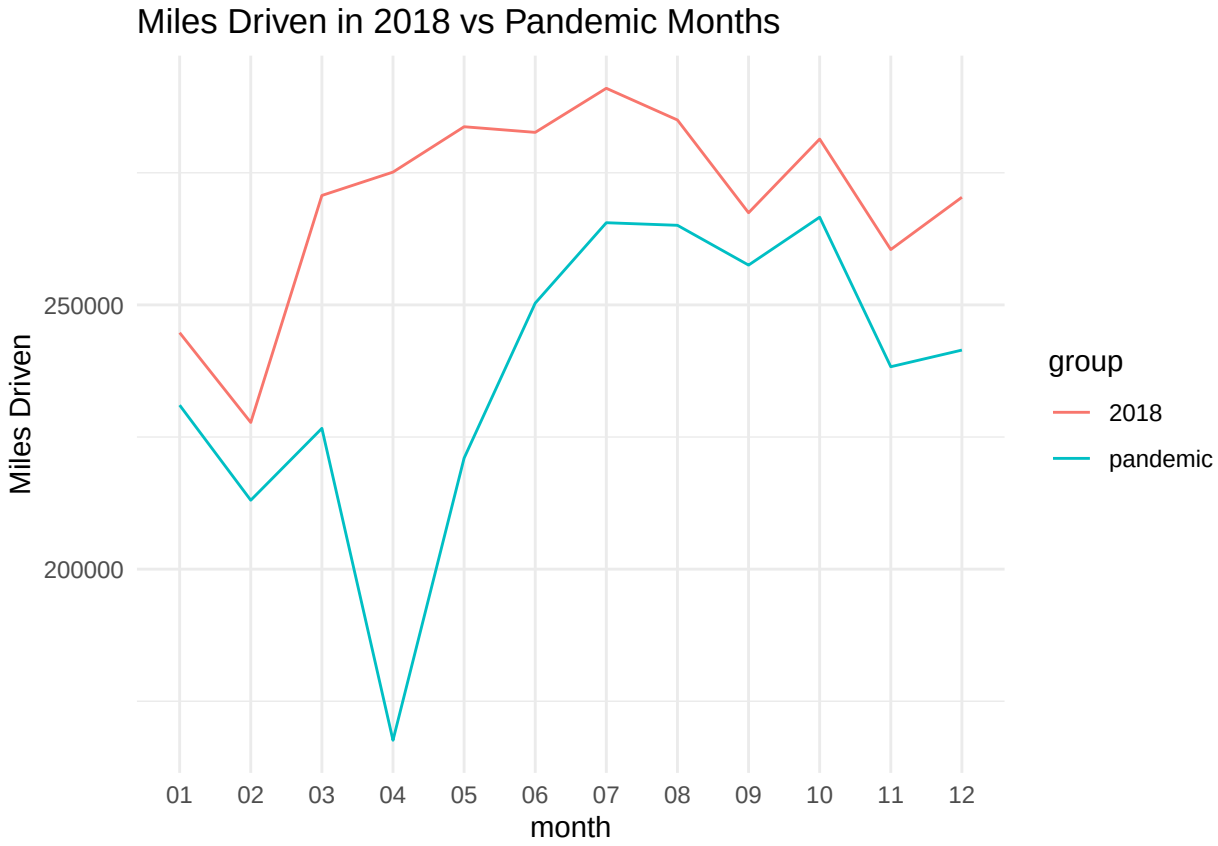
```
vehicle_miles$DATE <- as.Date(vehicle_miles$DATE)
vehicle_miles$year <- format(vehicle_miles$DATE, format="%Y")
head(vehicle_miles)
```

```
##         DATE TRFVOLUSM227NFWA year
## 1 1970-01-01             80173 1970
## 2 1970-02-01             77442 1970
## 3 1970-03-01             90223 1970
## 4 1970-04-01             89956 1970
## 5 1970-05-01             97972 1970
## 6 1970-06-01            100035 1970
```

We define pandemic months as the time frame between March 2020 (when COVID first hit) and March 2021 when the vaccine roll out had begun.

```
drive_2018 <- vehicle_miles %>% filter(year == 2018)
drive_pandemic <- vehicle_miles %>% filter(year == 2020 | year == 2021)
drive_2018$month <- format(drive_2018$DATE, "%m")
drive_pandemic$month <- format(drive_pandemic$DATE, "%m")
drive_pandemic <- drive_pandemic %>% slice(3:14)
drive_pandemic$group <- 'pandemic'
drive_2018$group <- 2018
comparison <- rbind(drive_2018, drive_pandemic)
```

```
ggplot(comparison, aes(x=month, y=TRFVOLUSM227NFWA, group=group)) +
  geom_line(aes(color=group)) + ylab('Miles Driven') +
  ggtitle('Miles Driven in 2018 vs Pandemic Months')
```



Visually from the plot above, we see the biggest differences coming in April and May. We will confirm further below.

```
drive_pandemic_month <- drive_pandemic %>% arrange(month)
differences <- drive_2018$TRFVOLUSM227NFWA -
  drive_pandemic_month$TRFVOLUSM227NFWA
perc <- scales::percent(drive_pandemic_month$TRFVOLUSM227NFWA[4] /
                          drive_2018$TRFVOLUSM227NFWA[4])
```

We find the months with the largest differences to have been April, 2018 had 1.0751×10^5 more million miles driven. In percentage terms, in April 2020, Americans drove 61% the amount that they did in April of 2018.

```
may_driving_2021 <- max(diff(drive_pandemic$TRFVOLUSM227NFWA))
perc2 <- scales::percent(drive_pandemic$TRFVOLUSM227NFWA[4] /
                           drive_pandemic$TRFVOLUSM227NFWA[3] - 1)
```

The maximum increase in driving was from April 2020 to May 2020, where driving increased by 5.3389×10^4 millions of miles. This represents a 13% increase month over month.

Now, use these changes in driving to make forecasts from the models.

```
effect1 <- between.model$coefficients['log(miles_driven_per_capita)'] *
  (drive_pandemic$TRFVOLUSM227NFWA[4]/drive_pandemic$TRFVOLUSM227NFWA[3])
```

If the number of miles driven per-capita increased by as much as the COVID boom, the conse-

quences of traffic fatalities would be expected to be an increase of 31.1369519 percent (in the log of the per-capita miles driven).

```
effect2 <- between.model$coefficients['log(miles_driven_per_capita)'] *
  (-(1-drive_pandemic_month$TRFVOLUSM227NFWA[4]/drive_2018$TRFVOLUSM227NFWA[4]))
```

If the number of miles driven per capita increased by as much as the COVID boom, the consequences of traffic fatalities would be expected to be an increase of -10.7419447 percent.

8 (5 points) Evaluate Error

If there were serial correlation or heteroskedasticity in the idiosyncratic errors of the model, what would be the consequences on the estimators and their standard errors? Is there any serial correlation or heteroskedasticity?

Serial correlation or heteroskedasticity in the idiosyncratic errors will cause the standard errors for the model to be incorrect. CLT can be invoked for first-difference models (Wooldridge). However, for fixed effects models, this is not necessarily the case as we'll likely see AR(1) effects in the error terms. Unless the AR(1) is a random walk we'll have to account for the AR(1) correlation. In case of random walk the first difference will be white noise and thus can be worked around. Let us use the Breusch-Godfrey test to check for serial correlation.

```
pbgtest(pool.model)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel models
##
## data:  total_fatality_rate ~ year + bac08 + bac10 + adm_lic_revoc_law +  ...
## chisq = 772.82, df = 25, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```
pbgtest(fd.model)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel models
##
## data:  total_fatality_rate ~ year + bac08 + bac10 + adm_lic_revoc_law +  ...
## chisq = 185.12, df = 25, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```
pbgtest(within.model)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel models
##
## data:  total_fatality_rate ~ year + bac08 + bac10 + adm_lic_revoc_law +  ...
## chisq = 324.4, df = 25, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

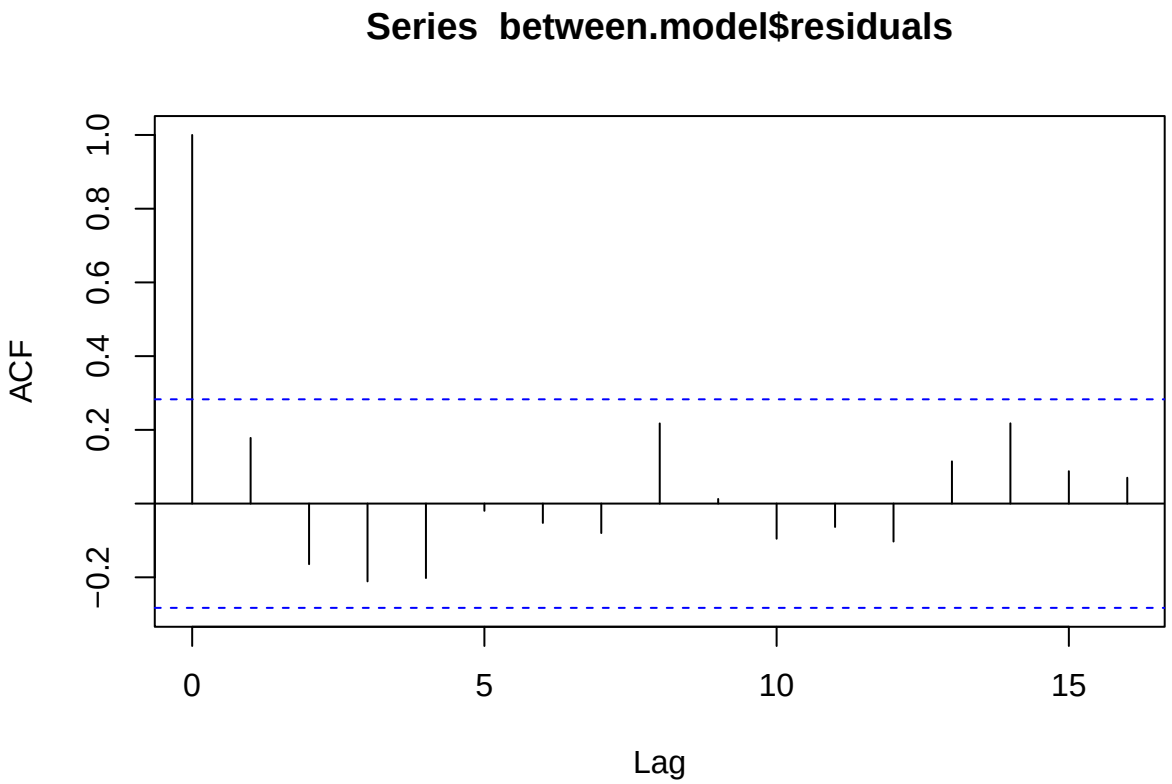
```
pbgtest(between.model)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel models
##
## data:  total_fatality_rate ~ year + bac08 + bac10 + adm_lic_revoc_law +  ...
## chisq = 30.613, df = 25, p-value = 0.2022
## alternative hypothesis: serial correlation in idiosyncratic errors
```

The test reveals for the serial correlation of the "between model" is the only one to not show serial correlation. The p-value does not reject the null hypothesis of no serial correlation. Thus the computed standard errors is likely to not have serial correlation.

Let us now look at the serial correlation of the residuals

```
acf(between.model$residuals)
```



The ACF plot confirms that the correlation is of significance only at lag 0, and thus no serial correlation.