Do changes in Traffic Laws affect traffic fatalities?

Srishti Mehra | David Djambazov

U.S. traffic fatalities: 1980-2004

1. Data Loading, Exploratory Data Analysis (EDA) on outcome *totfatrte* and the potential explanatory variables.

Data loading and review

```
load('driving.RData')
#str(data)
```

The structure of the data shows us that we have 1200 observations of data about traffic related laws, state attributes and year-wise indicator variables. The only value of state seen in the output above is 1 because there are 25 years of data we expect per state. Let's verify this by looking at the number of rows in this data per state.

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

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

This shows 25 rows of data per state. Let's validate what the number of rows per each year are:

```
table(data$year)
```

```
##
## 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
```

These are 48 per year. Each data point seems to be per state and per year because $25x48 = 1200$ which is our number of observations as seen in the structure of the data. We verified this by seeing how many rows exist per year per state.

We will now look at the first 5 rows of the dataset to see what the values for each column look like:

```
head(data,3)
```

```
##   year state sl55 sl65 sl70 sl75 slnone seatbelt minage zerotol gdl bac10 bac08
```

```
## 1 1980     1    1    0    0    0       0        0     18       0   0     1     0
## 2 1981     1    1    0    0    0       0        0     18       0   0     1     0
## 3 1982     1    1    0    0    0       0        0     18       0   0     1     0
##   perse totfat nghtfat wkndfat totfatpvm nghtfatpvm wkndfatpvm statepop
## 1     0    940     422     236      3.20      1.437      0.803  3893888
## 2     0    933     434     248      3.35      1.558      0.890  3918520
## 3     0    839     376     224      2.81      1.259      0.750  3925218
##   totfatrte nghtfatrte wkndfatrte vehicmiles unem perc14_24 sl70plus sbprim
## 1     24.14      10.84       6.06   29.37500  8.8      18.9        0      0
## 2     24.07      11.08       6.33   27.85200 10.7      18.7        0      0
## 3     21.37       9.58       5.71   29.85765 14.4      18.4        0      0
##   sbsecon d80 d81 d82 d83 d84 d85 d86 d87 d88 d89 d90 d91 d92 d93 d94 d95 d96
## 1       0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## 2       0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## 3       0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0
##   d97 d98 d99 d00 d01 d02 d03 d04 vehicmilespc
## 1   0   0   0   0   0   0   0   0     7543.874
## 2   0   0   0   0   0   0   0   0     7107.785
## 3   0   0   0   0   0   0   0   0     7606.622
```

In terms of understanding what the variables mean and what their units might be, we note that this dataset comes from the `wooldridge` package so we can find out more by running the following.

```
?driving
```

a. Speed limits: the variables `slXX` and '`slnone` are indicators for the speed limit in a given state in a given year. There's also a combined `sl70plus` variable indicating speed limits 70 mph and above.

b. Seatbelts: `seatbelt` is a categorical variable with values 0, 1, 2 corresponding to seatbelt not required, seatbelt enforcement is primary (that is a driver can be stopped and ticketed for a violation), seatbelt enforcement is secondary (seatbelt violation can be cited only if another primary violation is committed). There are also the indicator variables `sbprim` and `sbsecon`.

c. Drinking and driving: `minage` - minimum drinking age, `zerotol` - zero tolerance law, `bacXX` - indicator for blood alcohol limit, `perse` - administrative license revocation for DWI, `gdl` - graduated drivers license.

d. Fatality statistics: there are several variables for fatalities and conditions of occurrence as follows: in absolute numbers `totfat` - total, `nghtfat` - night time, `wkndfat` - weekend; in rate per 100 million vehicle miles as `xxxfatpvm`; in rate per 100K population as `xxxfatrte`.

e. Demographics: `unem` - unemployment rate in percent, `perc14_24` - percent of population aged 14-24, `vehicmiles` and `vehicmilespc` - vehicle miles per capita.

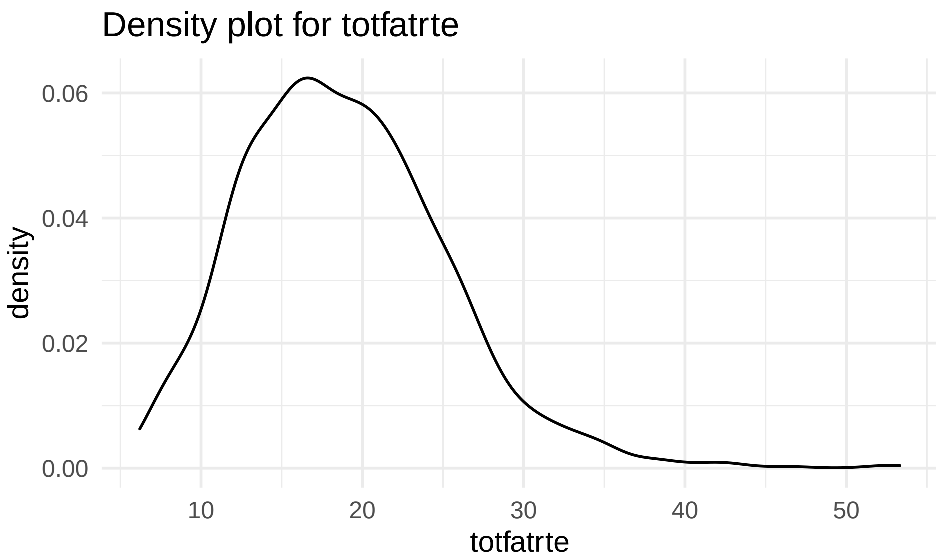
f. Year dummy variables indicating the year of observation.

EDA

Variable `totfatrte`

Density plot

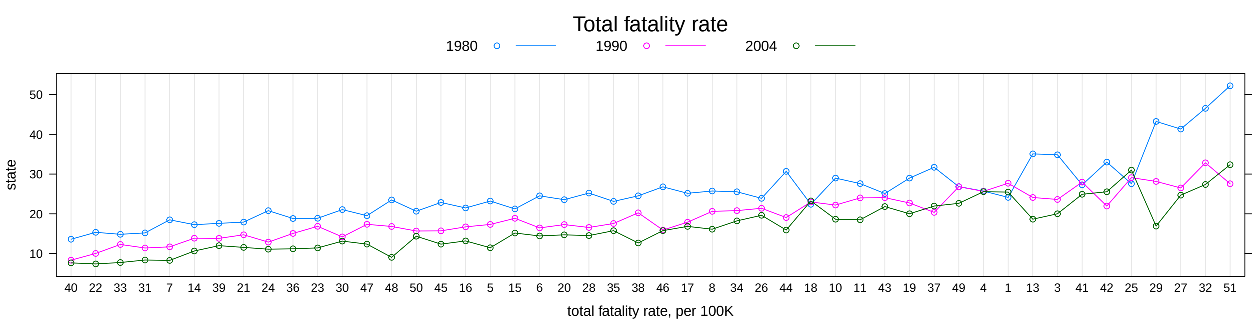
```
ggplot(data, aes(totfatrte)) + geom_density() + ggtitle("Density plot for totfatrte")
```



The density plot for Total Fatality per 100,000 population is a right skewed graph indicating that across years and states, lower fatality rates were more common in the range of all fatality rates. The most common fatality rate across years and states is close to 17 per 100,000 population.

Variance across states and panels. Let's look at `dotplot`.

```
dotplot(totfatrte ~ reorder(state, totfatrte),
        data[data$year %in% c(1980, 1990, 2004),],
        group=year,
        ylab="state", xlab="total fatality rate, per 100K",
        type=c("p", "a"),
        auto.key=list(columns=3, lines=TRUE, title="Total fatality rate"))
```

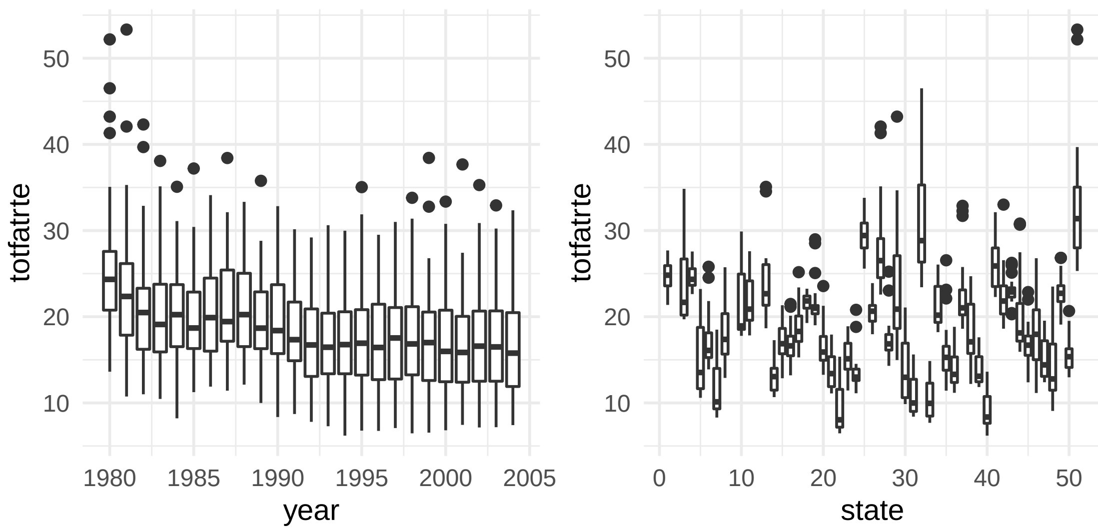


For better readability, we have chosen to only display data from 3 panels - the first in 1980, one in the middle (1990) and the final from 2004. We see that there's quite a bit a variability both across states and panels with 1980 having consistently higher values in almost all states. That give credence to the idea that some of the explanatory variables could have had an effect on decreasing the levels of `totfatrte` across the panels.

Boxplots of `totfatrte` Let's see the variance across states and panels.

```
p1 <- ggplot(data, aes(x=year, y=totfatrte, group = year)) +
  geom_boxplot()
p2 <- ggplot(data, aes(x=state, y=totfatrte, group = state)) +
  geom_boxplot()

p1 + p2
```



Confirming out intuition from the dotplot, we see that in the first half of the period there's a fairly consistent decrease in the total fatality rate across panels, followed by a period of relative stability of the mean. We also see some slight increase in the variance with time. The second boxplot indicates that the differences between states are quite significant, both in terms of mean and of variance.

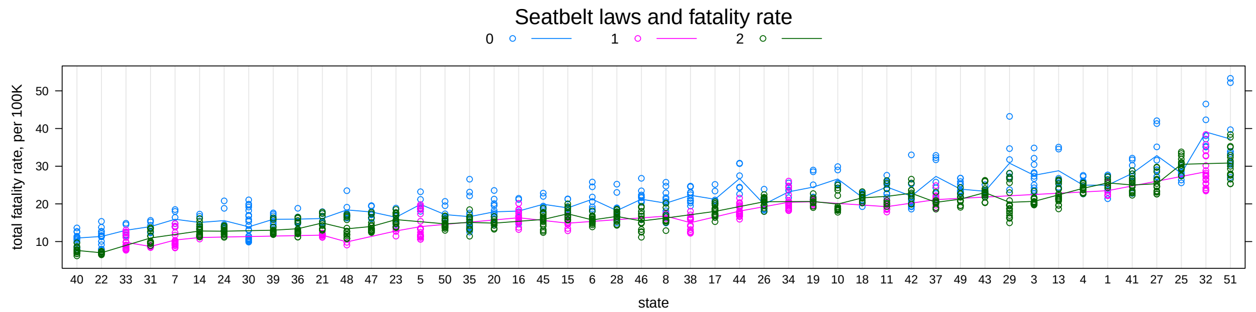
Explanatory variables

The variables that are of particular interest to us are those that could plausibly be connected to increase in traffic safety. So let's focus our attention on seatbelt laws, "per se" laws, speed and alcohol limits. In addition it's important to try to look if some demographic differences between states might help explain some of the variability.

Policy variables

a. Seatbelt laws

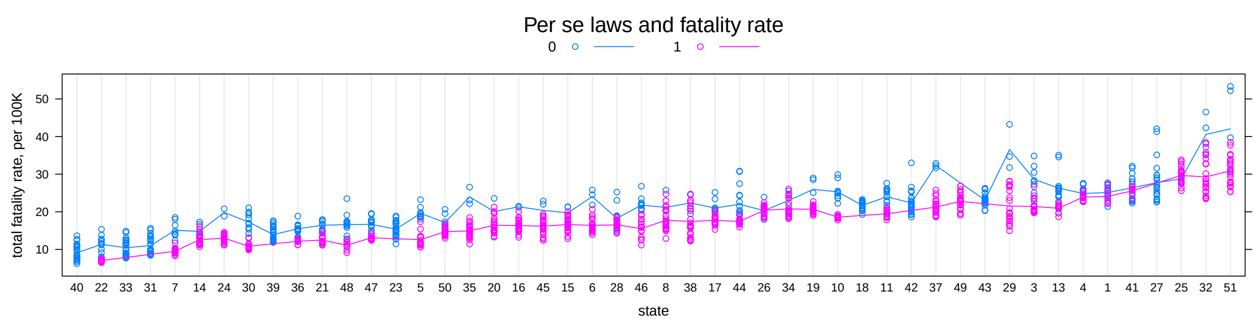
```
dotplot(totfatrte ~ reorder(state, totfatrte),
        data,
        group=seatbelt,
        xlab="state", ylab="total fatality rate, per 100K",
        type=c("p", "a"),
        auto.key=list(columns=3, lines=TRUE, title="Seatbelt laws and fatality rate"))
```



We see consistently across all states that the highest levels of `totfatrte` correspond to years in which there were no setabelt laws. Now that's not necessarily indicative of a causal relationship. One issue is that seatbelt laws were not introduced and retracted randomly, but are enacted at a given moment in time and all the latest observations are made under the seatbelt law. This can be problematic as we can imagine that if some other development with time is responsible for the decrease of fatalities (e.g. safer cars) even a non-existent effect might appear to be quite large. In fact, the one state that did not adopt a seatbelt law (30) also shows a similar trend. We do not know if people in that particular state began wearing seatbelts anyway as it became widely accepted and enforced in other states. Also interesting is that most states fall in either the primary or the secondary enforcement group. There's only a handful of states that have adopted both of the measures at some point and only a single state that has no seatbelt law as late as 2004.

b. "Per se" laws Let's look at the introduction of "per se" laws. Note that in order to account for the partial years without making the plot messier, we have rounded the variable to the nearest year. That should not materially affect the information displayed on the dotplot.

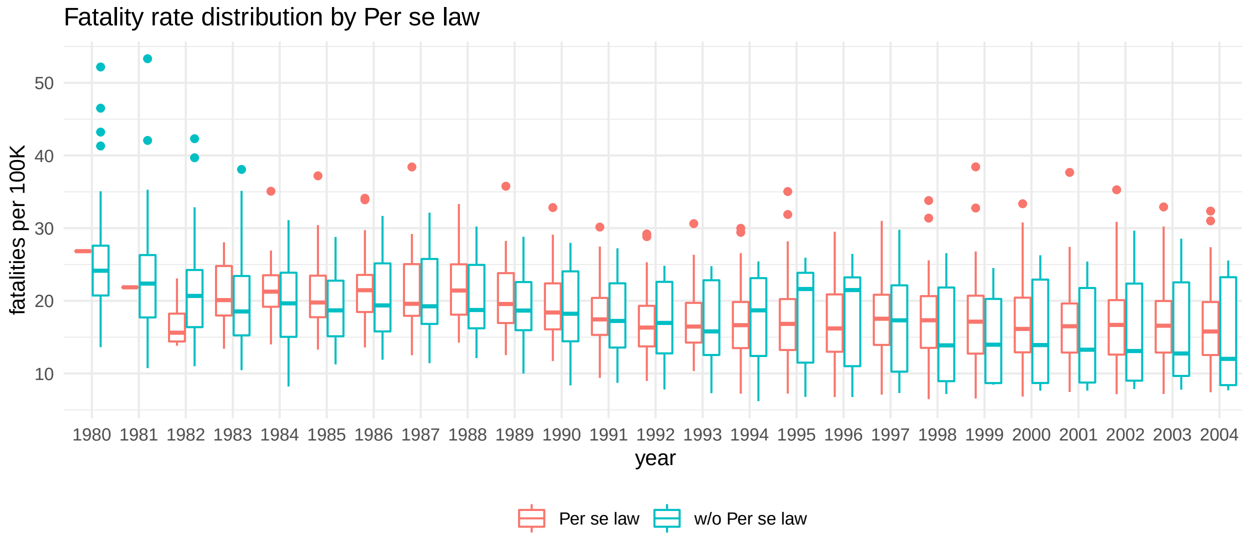
```
dotplot(totfatrte ~ reorder(state, totfatrte),
        data,
        group=as.factor(floor(perse+0.5)),
        xlab="state", ylab="total fatality rate, per 100K",
        type=c("p", "a"),
        auto.key=list(columns=2, lines=TRUE, title="Per se laws and fatality rate"))
```



Here again we see a similarly picture one concerning the seatbelt law introduction. One important feature we see is that even as fatality rates decrease and states adopt safety measure, with a few exceptions the relative ranking between states doesn't seem to change much.

```
ggplot(data, aes(as.factor(year), totfatrte)) +
  geom_boxplot(aes(color = as.factor(ifelse(floor(perse+0.5)>0,
                                            "Per se law",
                                            "w/o Per se law")))) +
```

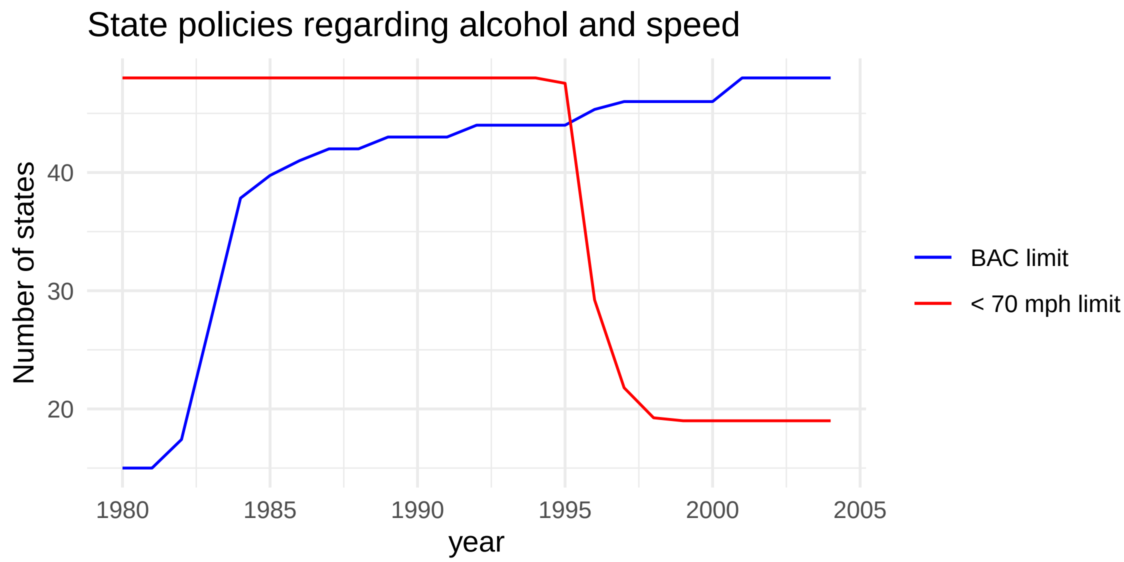
```
  theme(legend.title = element_blank(),
        legend.position = "bottom") +
  labs(title="Fatality rate distribution by Per se law",
       x = "year",
       y = "fatalities per 100K")
```



Looking at a by-year boxplot split between states with and without Per se laws, we see an interesting pattern. The sign of the difference in means between states with and without these laws seems to slowly oscillate in time. One possible explanation is that as the states with high fatalities adopt these laws, the mean of `totfatrte` for states with such laws increases, while the mean of states without them decreases. In the end we get mostly states with low fatalities in the latter category. There are other possible explanations, including that Per se laws do not affect fatality rates, while other factors do. So this does complicate making any visual determination if and how Per se laws affect the fatality rate.

c. Alcohol and speed limits

```
data %>%
  group_by(year) %>%
  summarise(bac = sum(bac08+bac10),
            sl70plus = 48 - sum(sl70plus)) %>%
  ggplot(aes(x=year)) +
  geom_line(aes(y=bac, col="blue")) +
  geom_line(aes(y=sl70plus, col="red")) +
  labs(title="State policies regarding alcohol and speed",
       y = "Number of states") +
  scale_color_manual(labels = c("BAC limit", "< 70 mph limit"),
                     values = c("blue", "red")) +
  theme(legend.title = element_blank())
```



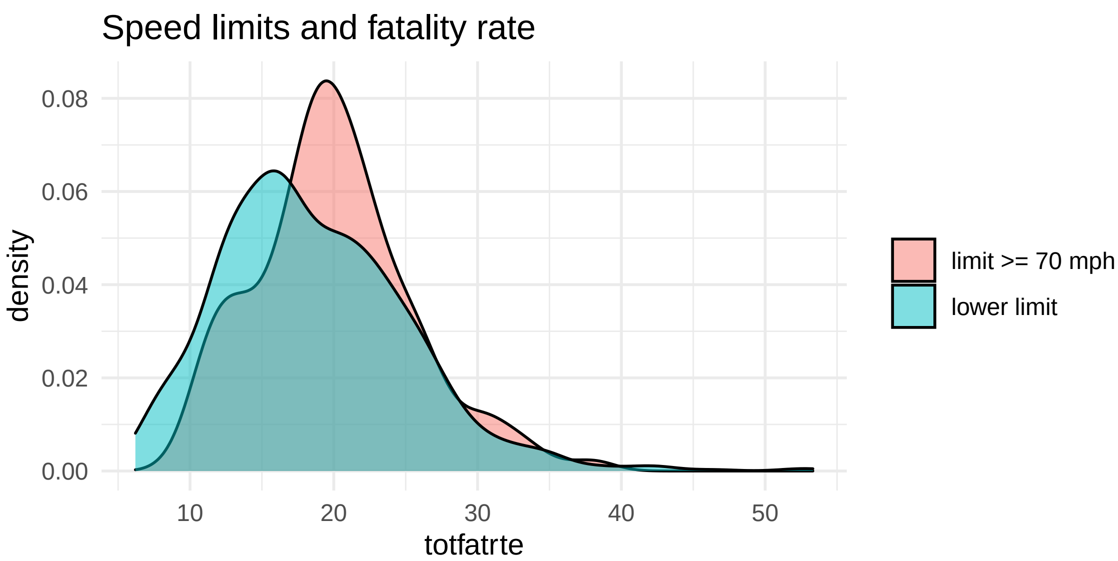

```
ggplot(data,aes(x=totfatrte, fill=as.factor(ifelse(bac08+bac10>0,"bac","no bac")))) +
  geom_density(alpha=0.5) +
  labs(title= "Blood Alcohol level limits and fatality rate") +
  theme(legend.title = element_blank())
```

The pooled distribution of fatality rates across year-states with BAC limits and those without, points to a potential effect of having those limits in place. Once again, we need to keep in mind that once BAC limits are adopted in a given state, they are not rescinded. Meaning that any improvement in the fatality rate with time (regardless of cause) will contribute to the differentiation the two densities as later values are measured under a BAC limit.

d. Speed limits

```
ggplot(data,aes(x=totfatrte, fill=as.factor(ifelse(floor(sl70plus+0.5)>0,
                                                    "limit >= 70 mph",
                                                    "lower limit")))) +
  geom_density(alpha=0.5) +
```

```
  labs(title= "Speed limits and fatality rate") +
  theme(legend.title = element_blank())
```



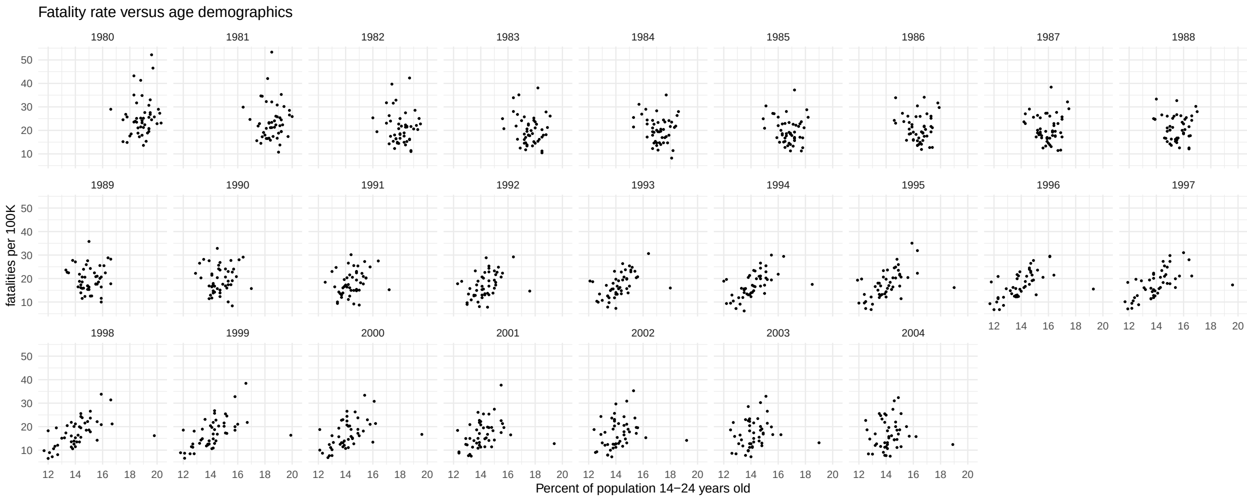
Here also see a potential influence of having lower speed limits on decreasing the fatality rate as the peak of the distribution is higher for the year-states with higher limits. Since states tend to increase speed limits with time, here we could argue somewhat more confidently that this might be a real effect.

Demographics Let's see some demographic factors that might offer insight into the differences between states.

a. Percentage of the population aged 14-24

Past studies have pointed to unsafe driving habits associated with younger drivers. So perhaps a larger prevelance of this age group in the population could be linked to a higher fatality rate. Let's take a look at scatter plots of `perc14-24` and `totfatrte` by year.

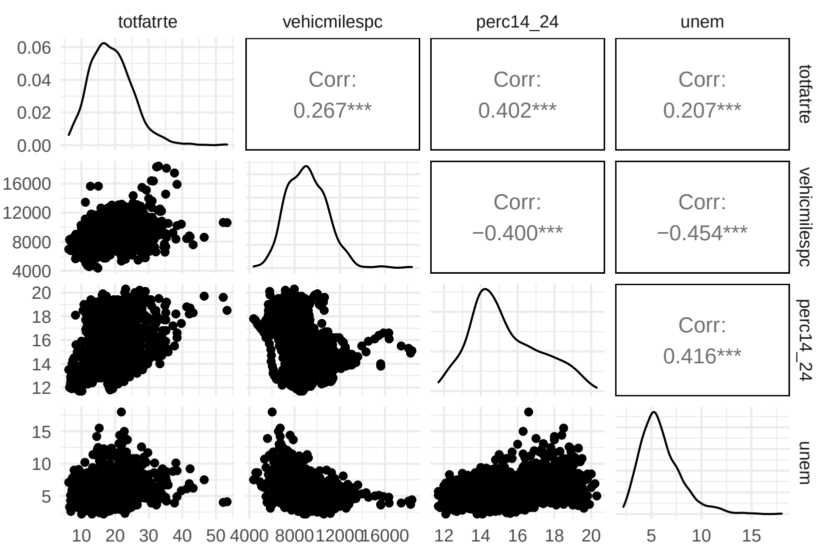
```
ggplot(data, aes(x=data$perc14_24,y=data$totfatrte)) +
  geom_point(size=0.5) +
  facet_wrap(~year,ncol=9) +
  labs(title = "Fatality rate versus age demographics",
       y = "fatalities per 100K",
       x = "Percent of population 14-24 years old")
```



This graph paints a story of two periods. Prior to 1992 we don't see any indication of a linear relationship between the two variables in any given year. We do see, however, an overall movement of the datapoints to the lower left as the fatality rate decreases together with the proportion of younger people. Starting in 1992 we also see the data stretch out in a more linear pattern with states with higher proportion of young people featuring a higher fatality rate. In the 2000s we again we again see the relationship weaken.

b. Correlations between demographic variables

```
ggpairs(data[c("totfatrte","vehicmilespc","perc14_24","unem")])
```



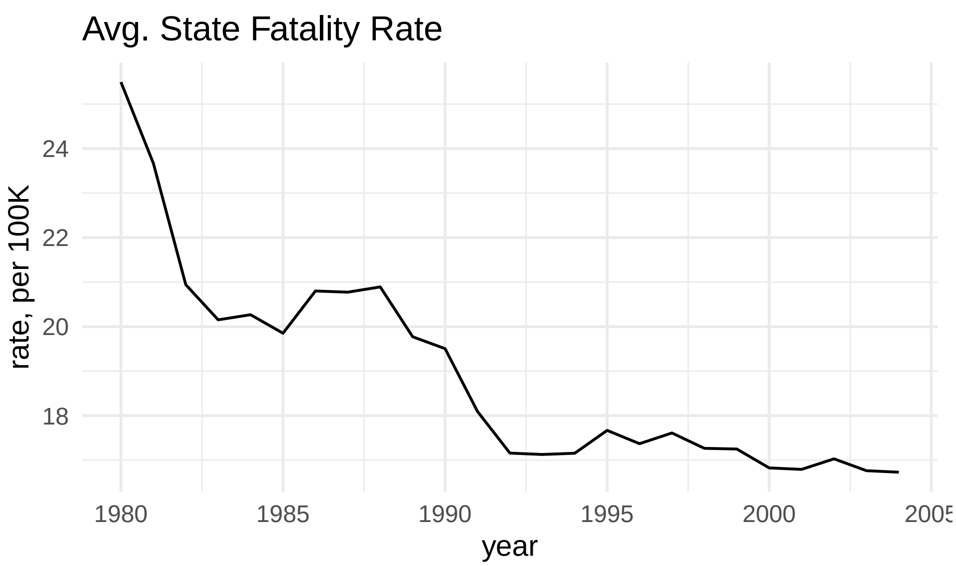
From the correlation plot above we can make several important observations. Firstly, our response variable `totfarte` as well as `vehicmilespc` and `unem` seem right skewed. Secondly, there are some correlations, but they are not too pronounced. Peredominantly we see a confirmation of the above two plots where we saw some relationship between `totfatrte` on one hand and `perc14_24` and `vehicmilespc` on the other. Unemployment does not seem to be much of a direct factor. However we do see a negative correlation between unemployment vehicle miles traveled per capita. As with many of the variables there could easily be a confounding variable - as the economy has grown with time, unemployment has decreased, while miles traveled have increased. At any rate, it would be interesting to include all these variables into our models.

2. Definition of outcome variable *totfatrte*, average of this variable in each of the years in the time period covered in this dataset; A linear regression model of *totfatrte* on a set of dummy variables for the years 1981 through 2004 to see on high level if driving become safer over this period.

`totfatrte` is defined as fatalities per 100K population. That is, `totfatrte` should equal the ratio between `totfat` and `statepop` for each state. If we want to calculate an average total fatality rate for a given year there are a couple of approaches. One would be to take a simple arithmetic average of the value for each state. The second would be to sum up all the fatalities and state populations for a given year and take the ratio. In terms of our statistical analysis, the former would be more appropriate. For our purposes we consider each state a unit of observation. There's a question if it is justified to consider states independent units, but acknowledging that possible challenge we can try to make sense of the mean of `totfatrte` across units.

```
totfatrte.mean <- data %>%
  group_by(year) %>%
  summarise(mean = mean(totfatrte))

ggplot(totfatrte.mean, aes(x=year, y=mean)) +
  geom_line() +
  #ylim(0,30) +
  labs(title="Avg. State Fatality Rate",
       y = "rate, per 100K")
```



Let's run the model.

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

```
##
```

```
## Call:
## lm(formula = totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 +
##     d87 + d88 + d89 + d90 + d91 + d92 + d93 + d94 + d95 + d96 +
##     d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9302  -4.3468  -0.7305   3.7488  29.6498
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  25.4946     0.8671  29.401  < 2e-16 ***
## 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
```

The model takes 1980 as the base case, with the intercept corresponding to the mean of `totfatrte` for that year. Each coefficient then is literally the difference between the base case (1980) and the corresponding year's mean. What we find is that all the coefficients are negative, so each year's mean fatality rate is less than that in1980. We also see that as we go from 1980 to 2004, the coefficients mostly increase in absolute value. In fact the d04 coefficient is the most negative of all. Also all coefficients with the exception of d81 are statistically significant. From this we can say that there is evidence to reject the null hypothesis that the average `totfatrte` per state has remain unchanged versus 1980 for all years but 1981. We still need to be a bit careful answering the question if driving became safer over this period. On the basis of our analysis we can say that the fatality rate per 100K in 2004 was quite a bit less than that in 1980 for any randomly chosen state. If the question is whether the trend over that period was always negative, then we cannot assertively answer. Also, as we noted before, here we're treating each state as an individual unit. Since safety of driving is determined by the probability of a person being in a fatal accident, not a state, we might need to also weigh for the population as a whole. If a large state became less safe, the mean across states could go down, while the fatality rate for all states combined could go up.

3. Expanding model from *Part 2* by adding variables *bac08, bac10, perse, sbprim, sbsecon, sl70plus, gdl, perc14_24, unem, vehicmilespc*, and perhaps *transformations of some or all of these variables.*

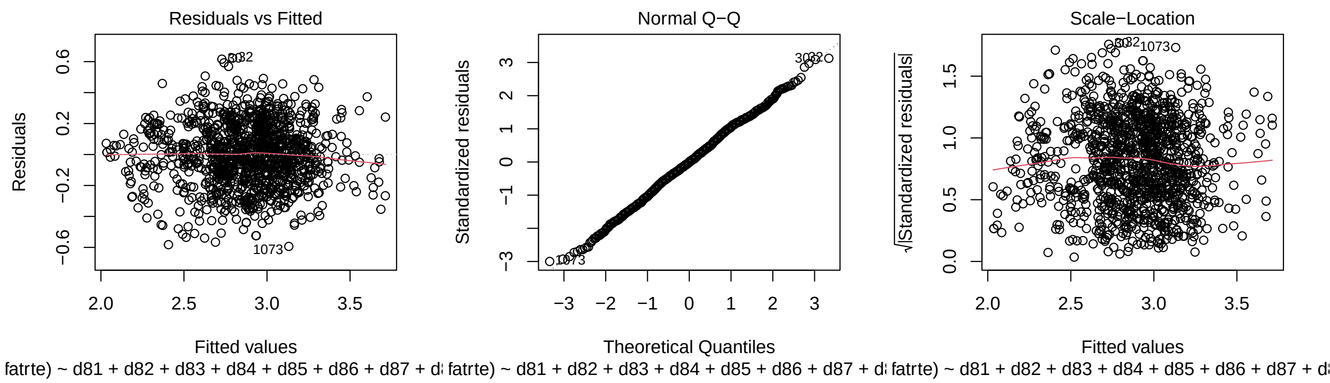
We did log transformation for *unem*, *vehicmilespc*, and *totfatrte* because of their right-skewness (higher lower values) as seen in the ggpairs plot above. We also saw that variance in `totfatrte` has increased with time. A log transformation would also help us with the effect heteroskedascity of errors could have on our model.

```
lin.reg2 <- lm(log(totfatrte) ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 +
                d89 + d90 + d91 + d92 + d93 + d94 + d95 + d96 + d97 + d98 +
                d99 + d00 + d01 + d02 + d03 + d04 +
                floor(bac08+0.5) + floor(bac08+0.5) + floor(bac10+0.5) +
                floor(perse+0.5) + floor(sbprim+0.5) + floor(sbsecon+0.5) +
                floor(sl70plus+0.5) + floor(gdl+0.5) +
                perc14_24 + log(unem) + log(vehicmilespc), data=data)
summary(lin.reg2)
```

```
##
## Call:
## lm(formula = log(totfatrte) ~ d81 + d82 + d83 + d84 + d85 + d86 +
##     d87 + d88 + d89 + d90 + d91 + d92 + d93 + d94 + d95 + d96 +
##     d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04 + floor(bac08 +
##     0.5) + floor(bac08 + 0.5) + floor(bac10 + 0.5) + floor(perse +
##     0.5) + floor(sbprim + 0.5) + floor(sbsecon + 0.5) + floor(sl70plus +
##     0.5) + floor(gdl + 0.5) + perc14_24 + log(unem) + log(vehicmilespc),
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59354 -0.12705 -0.00089  0.13981  0.62265
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -1.132e+01  4.023e-01 -28.147  < 2e-16 ***
## d81                    -9.180e-02  4.124e-02  -2.226  0.02621 *
## d82                    -2.946e-01  4.211e-02  -6.995 4.46e-12 ***
## d83                    -3.495e-01  4.298e-02  -8.133 1.06e-15 ***
```

```
## d84                  -2.993e-01  4.371e-02  -6.846 1.22e-11 ***
## d85                  -3.373e-01  4.460e-02  -7.563 7.98e-14 ***
## d86                  -3.142e-01  4.644e-02  -6.765 2.10e-11 ***
## d87                  -3.500e-01  4.842e-02  -7.229 8.81e-13 ***
## d88                  -3.602e-01  5.096e-02  -7.068 2.70e-12 ***
## d89                  -4.454e-01  5.292e-02  -8.417  < 2e-16 ***
## d90                  -5.050e-01  5.410e-02  -9.334  < 2e-16 ***
## d91                  -6.193e-01  5.531e-02 -11.197  < 2e-16 ***
## d92                  -7.265e-01  5.632e-02 -12.899  < 2e-16 ***
## d93                  -7.170e-01  5.708e-02 -12.562  < 2e-16 ***
## d94                  -7.027e-01  5.822e-02 -12.071  < 2e-16 ***
## d95                  -6.815e-01  5.949e-02 -11.457  < 2e-16 ***
## d96                  -8.141e-01  6.187e-02 -13.157  < 2e-16 ***
## d97                  -8.149e-01  6.304e-02 -12.927  < 2e-16 ***
## d98                  -8.607e-01  6.383e-02 -13.485  < 2e-16 ***
## d99                  -8.595e-01  6.488e-02 -13.247  < 2e-16 ***
## d00                  -8.713e-01  6.594e-02 -13.213  < 2e-16 ***
## d01                  -9.215e-01  6.690e-02 -13.775  < 2e-16 ***
## d02                  -9.652e-01  6.727e-02 -14.350  < 2e-16 ***
## d03                  -9.875e-01  6.757e-02 -14.615  < 2e-16 ***
## d04                  -9.733e-01  6.886e-02 -14.134  < 2e-16 ***
## floor(bac08 + 0.5)    -5.648e-02  2.446e-02  -2.310  0.02109 *
## floor(bac10 + 0.5)    -1.215e-02  1.805e-02  -0.673  0.50109
## floor(perse + 0.5)    -2.073e-02  1.453e-02  -1.426  0.15412
## floor(sbprim + 0.5)    9.265e-04  2.462e-02   0.038  0.96999
## floor(sbsecon + 0.5)   2.089e-02  2.145e-02   0.974  0.33043
## floor(sl70plus + 0.5)  2.175e-01  2.161e-02  10.062  < 2e-16 ***
## floor(gdl + 0.5)      -2.789e-02  2.512e-02  -1.110  0.26718
## perc14_24              1.833e-02  6.112e-03   2.999  0.00276 **
## log(unem)              2.684e-01  2.415e-02  11.111  < 2e-16 ***
## log(vehicmilespc)      1.543e+00  4.445e-02  34.707  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2017 on 1165 degrees of freedom
## Multiple R-squared:  0.6672, Adjusted R-squared:  0.6575
## F-statistic:  68.7 on 34 and 1165 DF,  p-value: < 2.2e-16
```

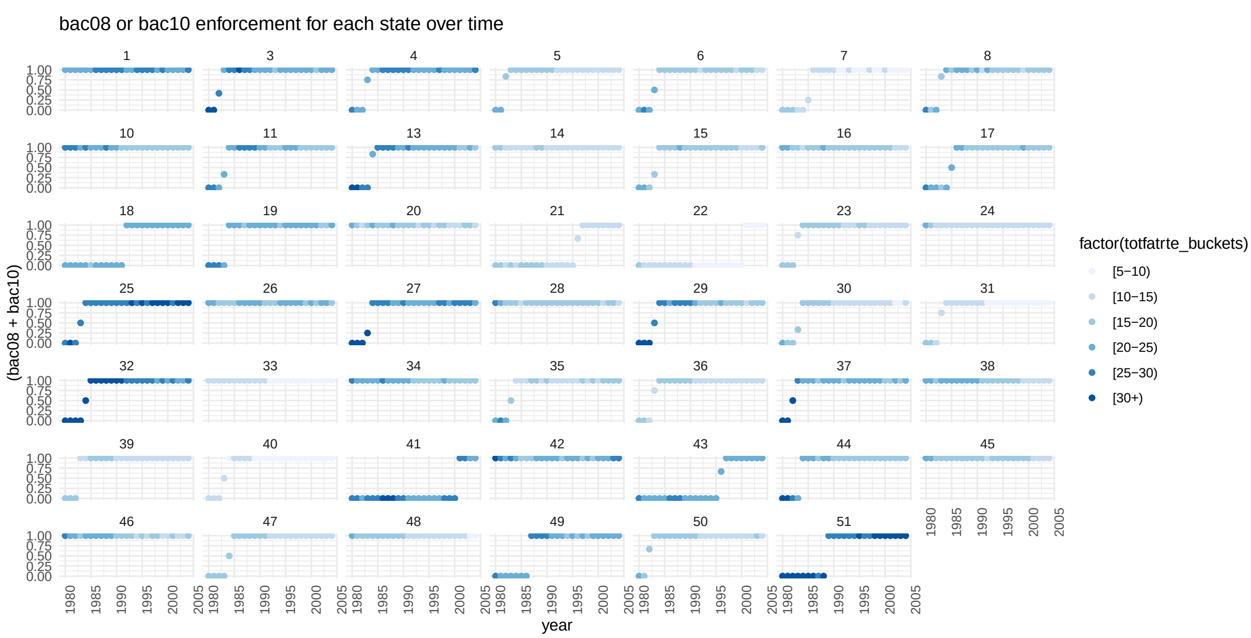
```r
plot(lin.reg2, which=1:3)
```



From the residual plots we see that the residuals of the model are well behaved. The QQ plot and the plot versus fitted values indicate that the variable transformations we made were quite reasonable.

```r
breaks <- c(0,5,10,15,20,25,30,100)
tags <- c("[0-5)","[5-10)", "[10-15)", "[15-20)", "[20-25)", "[25-30)","[30+)")

data <- data %>% mutate(totfatrte_buckets = cut(data$totfatrte,
                                                 breaks=breaks,
                                                 include.lowest=TRUE,
                                                 right=FALSE, labels=tags))
library(RColorBrewer)
p <-  ggplot(data = data, aes(x = year, y = (bac08+bac10),
                               color=factor(totfatrte_buckets))) +
  geom_point() +
  facet_wrap(~state)
p + labs(title="bac08 or bac10 enforcement for each state over time") +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_color_brewer(palette = "Blues")
```



bac08 and *bac10* are variables that indicate (1 or 0) whether or not there is a blood alcohol limit of .08 or 0.10 in the state. We can see from the graphs above that these limits have changed over time in states. Of these two, *bac08* is a statistically significant explanatory variable for *totfatrte*. The coefficient of *bac08* in the linear regression tells us that by increasing *bac08* by 1 unit (enforcing the blood alcohol limit of .08) we can change *totfatrte* by $(exp(-0.05648)-1) * 100 =$ -5.49%.

perse laws have a negative effect on *totfatrte*, however, it is not a statistically significant explanatory variable that if increases by 1 unit would change *totfatrte* by $(exp(-0.02073)-1) * 100 =$ -2%.

Having a primary seat belt law, indicated by *sbprim*, also does not have a statistically significant effect on *totfatrte*. It's coefficient indicates that increasing *sbprim* by 1 unit (enforcing a primary seat belt law) would change *totfatrte* by $(exp(0.0009265)-1) * 100 =$ +0.09%.

4. Reestimating the model from *Part 3* using a fixed effects (at the state level) model.

```
data.panel <- pdata.frame(data, index=c("state", "year"))

model.plm <- plm(log(totfatrte) ~ year +
                floor(bac08+0.5) + floor(bac08+0.5) + floor(bac10+0.5) +
                floor(perse+0.5) + floor(sbprim+0.5) + floor(sbsecon+0.5) +
                floor(sl70plus+0.5) + floor(gdl+0.5) +
                perc14_24 + log(unem) + log(vehicmilespc), data=data.panel, model="within")

summary(model.plm)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log(totfatrte) ~ year + floor(bac08 + 0.5) + floor(bac08 +
##     0.5) + floor(bac10 + 0.5) + floor(perse + 0.5) + floor(sbprim +
##     0.5) + floor(sbsecon + 0.5) + floor(sl70plus + 0.5) + floor(gdl +
##     0.5) + perc14_24 + log(unem) + log(vehicmilespc), data = data.panel,
##     model = "within")
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -0.3801338 -0.0511013  0.0041506  0.0530428  0.2885775
##
## Coefficients:
##                         Estimate Std. Error  t-value  Pr(>|t|)
## year1981              -0.0631255  0.0180650  -3.4944 0.0004938 ***
## year1982              -0.1349227  0.0189854  -7.1067 2.114e-12 ***
## year1983              -0.1691547  0.0197174  -8.5790 < 2.2e-16 ***
## year1984              -0.2093631  0.0204594 -10.2331 < 2.2e-16 ***
## year1985              -0.2347146  0.0214083 -10.9637 < 2.2e-16 ***
## year1986              -0.1981537  0.0229486  -8.6347 < 2.2e-16 ***
## year1987              -0.2443068  0.0249302  -9.7996 < 2.2e-16 ***
## year1988              -0.2748302  0.0273099 -10.0634 < 2.2e-16 ***
```

```
## year1989              -0.3493713  0.0291264 -11.9950 < 2.2e-16 ***
## year1990              -0.3591244  0.0302907 -11.8559 < 2.2e-16 ***
## year1991              -0.3958058  0.0310302 -12.7555 < 2.2e-16 ***
## year1992              -0.4563916  0.0320456 -14.2420 < 2.2e-16 ***
## year1993              -0.4742549  0.0326815 -14.5114 < 2.2e-16 ***
## year1994              -0.5067797  0.0335977 -15.0838 < 2.2e-16 ***
## year1995              -0.5072995  0.0345822 -14.6694 < 2.2e-16 ***
## year1996              -0.5627387  0.0367652 -15.3063 < 2.2e-16 ***
## year1997              -0.5832943  0.0377717 -15.4426 < 2.2e-16 ***
## year1998              -0.6367058  0.0385257 -16.5268 < 2.2e-16 ***
## year1999              -0.6538255  0.0390750 -16.7326 < 2.2e-16 ***
## year2000              -0.6863511  0.0396573 -17.3071 < 2.2e-16 ***
## year2001              -0.6555035  0.0400635 -16.3616 < 2.2e-16 ***
## year2002              -0.6178539  0.0403357 -15.3178 < 2.2e-16 ***
## year2003              -0.6211582  0.0405671 -15.3119 < 2.2e-16 ***
## year2004              -0.6603615  0.0414140 -15.9454 < 2.2e-16 ***
## floor(bac08 + 0.5)    -0.0049901  0.0143324  -0.3482 0.7277769
## floor(bac10 + 0.5)    -0.0056711  0.0099929  -0.5675 0.5704813
## floor(perse + 0.5)    -0.0564198  0.0098031  -5.7553 1.116e-08 ***
## floor(sbprim + 0.5)   -0.0405369  0.0149934  -2.7037 0.0069621 **
## floor(sbsecon + 0.5)   0.0059959  0.0110063   0.5448 0.5860178
## floor(sl70plus + 0.5)  0.0728376  0.0113768   6.4023 2.248e-10 ***
## floor(gdl + 0.5)      -0.0222396  0.0122103  -1.8214 0.0688173 .
## perc14_24              0.0197779  0.0041609   4.7533 2.262e-06 ***
## log(unem)             -0.1930972  0.0171792 -11.2402 < 2.2e-16 ***
## log(vehicmilespc)      0.6761190  0.0508574  13.2944 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    31.924
## Residual Sum of Squares: 8.682
## R-Squared:      0.72804
## Adj. R-Squared: 0.70834
## F-statistic: 88.0282 on 34 and 1118 DF, p-value: < 2.22e-16
```

Differences between the results from fitting a fixed effects model vs. fitting a pooled OLS model:

bac08 is not a statistically significant explanatory variable for *totfatrte* when evaluated with a fixed effects model, unlike when evaluated with a pooled OLS model. In terms of the effect, a one unit increase in *bac08* brings a $(exp(-0.0049901) - 1) * 100 =$ -0.5% change in *totfatrte*.

bac10 is now a slightly statistically significant explanatory variable for *totfatrte* when evaluated with a fixed effects model, unlike when evaluated with a pooled OLS model. For one unit increase in *bac10*, there is a $(exp(-0.0056711) - 1) * 100 =$ -0.56% change in *totfatrte*.

perse is now a highly statistically significant explanatory variable for *totfatrte* when evaluated with a fixed effects model, unlike when evaluated with a pooled OLS model. For one unit increase in *perse*, there is a $(exp(-0.0564198) - 1) * 100 =$ -5.5% change in *totfatrte*.

sbprim is now a statistically significant explanatory variable for *totfatrte* when evaluated with a fixed effects model, unlike when evaluated with a pooled OLS model. For one unit increase in *sbprim*, there is a $(exp(-0.0405369) - 1) * 100 =$ -4% change in *totfatrte*.

We think the estimates from the fixed effects models are more reliable because the assumptions for the fixed effects model are more plausible to hold as compared to the assumptions for the pooled OLS model. Assumptions:

Fixed Effects model - The idiosyncratic error, that varies with state and year, should be uncorrelated with each explanatory variable across all time periods. It allows for arbitrary correlation between time invariant unobserved effects and the explanatory variables. If we had a time-invariant unobserved variable like natural resources in each state which would effect an explanatory variable like unemployment, then we would be safe using Fixed Effects model. We, in the fixed effects model, essentially run a pooled OLS after subtracting the demeaned panel data with the actual panel data. In that process, we eliminate the time-invariant effects (observed and unobserved).

Pooled OLS Model - The pooled OLS requires that the composite error term consisting of both the time-variant and time-invariant unobserved effects is uncorrelated with the explanatory variables. A possible violation for this assumption could be the unobserved effect "technological advance" affecting our response *totfatrte* as it could also be correlated to the *unem* explanatory variable in our model. Since this is a more restricting assumption, it is less likely to be met.

5. Evaluating use of a random effects model instead of the fixed effects model built in *Part 4*.

A random effects model requires the assumption that the unobserved effects are uncorrelated with each explanatory variable to hold. This assumption is hard to meet as mentioned above with an example in part 4 and in general with the interactions that our explanatory variables would have with law-enforcement and econometrics factors that are not observed. With a restrictive assumption and the main benefit of the random effects model being to be able to estimate effect of time-invariant variables on the response variable, we would not prefer to use the random effects model. All the variables currently in our interest as explanatory variables for *totfatrte* are time-variant so we will not get a lot of benefit from using the random effects model.

6. If *vehicmilespc*, the number of miles driven per capita, increases by $1,000$, what would be the estimated effect on *totfatrte*?

The variable *vehicmilespc* has a highly statistically significant effect on *totfatrte*, as seen from our fixed effects model. Since we did a log tansformation for the variable while adding it to the model, we will need to estimate the change percent associated with a $1,000$ gain in miles.

```
mean(data$vehicmilespc)
```

```
## [1] 9129.044
```

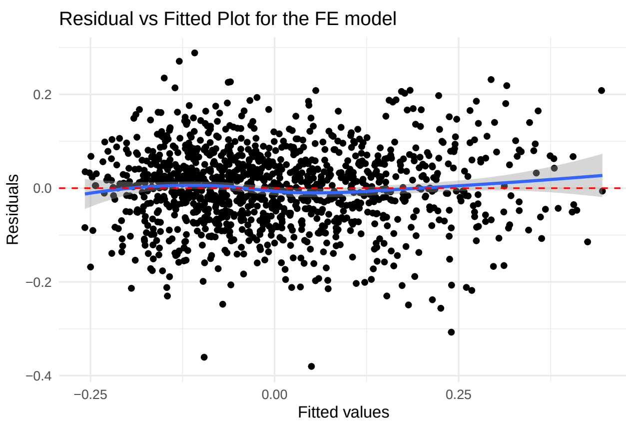
From the mean of 9129, a gain of 1,000 is $\frac{1000}{9129} * 100$=10.9%. The coefficient in the model tells us that for a 10.9% gain from baseline 9129 *vehicmilespc* we estimate to see a $(1.109^{0.6771286}-1) * 100$ = 7.25% change in *totfatrte*. In general, depending on the baseline *vehicmilespc* value being chosen, if the % change that a gain of 1,000 causes is x, then the estimated change in *totfatrte* is $((1 + \frac{x}{100})^{0.6771286} - 1) * 100$

7. Evaluating serial correlation or heteroskedasticity in the idiosyncratic errors of the model, and their consequences on the estimators and their standard errors.

```r
plm.res.fit.df <- data.frame(res = residuals(model.plm), fit = fitted(model.plm))
```

```r
ggplot(plm.res.fit.df, aes(fit, res)) + geom_point() +
  stat_smooth(method="loess") + geom_hline(yintercept=0, col="red", linetype="dashed") +
  xlab("Fitted values")+ylab("Residuals") +
  ggtitle("Residual vs Fitted Plot for the FE model")
```

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



The residuals vs. fitted values plot shows us that there is no heteroskedasticity.

```r
bgtest(model.plm, order = 1, order.by = NULL, type = c("Chisq"))
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  model.plm
## LM test = 800.61, df = 1, p-value < 2.2e-16
```

The Breusch-Godfrey test shows evidence to reject the null hypothesis that there is no serial correlation of any order up to 1. This indicates that we may have inefficient estimates and biased standard errors. A better model may be one that is able to take the existing auto correlations into account.