# W271: Lab 4

*Morris Burkhardt/group*

*April 22, 2018*

```r
library(ggplot2); library(dplyr); library(lme4); library(Hmisc); library(plm); library(gridExtra)
```

## Question 1.

```r
load('driving.RData')
head(data, 2); #tail(data); #summary(data); #str(data)
```

```
##   year state sl55 sl65 sl70 sl75 slnone seatbelt minage zerotol gdl bac10
## 1 1980     1    1    0    0    0      0        0     18       0   0     1
## 2 1981     1    1    0    0    0      0        0     18       0   0     1
##   bac08 perse totfat nghtfat wkndfat totfatpvm nghtfatpvm wkndfatpvm
## 1     0     0    940     422     236      3.20      1.437      0.803
## 2     0     0    933     434     248      3.35      1.558      0.890
##   statepop totfatrte nghtfatrte wkndfatrte vehicmiles unem perc14_24
## 1  3893888     24.14      10.84       6.06     29.375  8.8      18.9
## 2  3918520     24.07      11.08       6.33     27.852 10.7      18.7
##   sl70plus sbprim sbsecon d80 d81 d82 d83 d84 d85 d86 d87 d88 d89 d90 d91
## 1        0      0       0   1   0   0   0   0   0   0   0   0   0   0   0
## 2        0      0       0   0   1   0   0   0   0   0   0   0   0   0   0
##   d92 d93 d94 d95 d96 d97 d98 d99 d00 d01 d02 d03 d04 vehicmilespc
## 1   0   0   0   0   0   0   0   0   0   0   0   0   0     7543.874
## 2   0   0   0   0   0   0   0   0   0   0   0   0   0     7107.785
```

First, we looked at the head, tail, summary and str of the data to get an idea of the data set (partially not displayed here).

The given data set is a longitudinal data set. It consists of the data of the 48 continental states over the time period from 1980 to 2004.

The data set has the following 56 variables:

- year = year of observation; integer; ranges from 1980 to 2004
- state = state; integer; ranges from 1 to 51 (missing 2, 9 and 12: Alaska, Hawai and D.C.)
- sl55, sl65, sl70, sl75, slnone = speed limit 55, 65, 70, 75 and non respectively; decimal; fraction of year; all these five variables sum up to 1
- seatbelt = indicates what type of seatbelt law was inplace: no seatbelt law (=0), primary seatbelt law (=1), secondary seatbelt law (=2); integer; ranges from 0 to 2
- minage = minimum drinking age; decimal; weighted yearly minimum drinking age; ranges from 18 to 21
- zerotol = zero tolerance law; decimal; fraction of year for which a zero tolerance was in place
- gdl = graduated driver license law; decimal; fraction of year for which a gdl law was in place
- bac10, bac08 = blood alcohol limit of 0.1, 0.08 respectively; decimal; fraction of year for which each limit was in place
- perse = per se law; decimal; fraction of year for which a per se law was in place
- totfat, nghtfat, wkndfat = integer
- totfatpvm, nghtfatpvm, wkndfatpvm = total, nighttime, weekend fatalities, respectively; integer
- statepop = state population; integer
- totfatrte, nghtfatrte, wkndfatrte = total, nighttime and weekend fatalities per 100,000 population, respectively; decimal
- vehicmiles = billion vehicle miles traveled; decimal

- unem = unemployment rate in percent; decimal; ranges from 2.2 to 18
- per14_24 = percentage of population aged 14 through 24; decimal; ranges from 11.7 to 20.3
- sl70plus = sum of the sl70, sl75 and slnone variables; decimal; fraction of year
- sbprim, sbsecon = primary, secondary seatbelt law respectively; dummy encoding of seatbelt variable
- d80, d81, . . . , d04 = year 1980, year 1981,. . . year 2004 respectively; dummy encoding of year variable
- vehiclemilespc = vehicle miles per capita; decimal

Let us check if our panel is balanced:

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

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

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

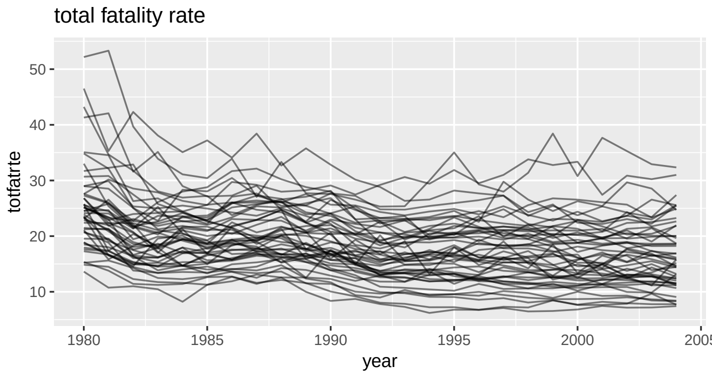
Each of the 48 states has exactly 25 observations across time, which means that our panel is balanced. Next, let us check for missing data.

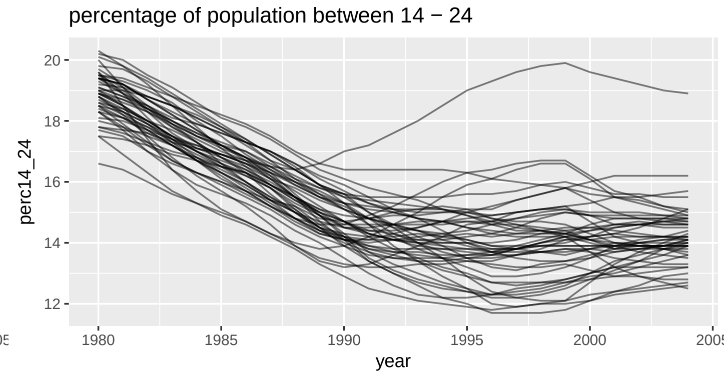
```
sum(is.na(data))
```

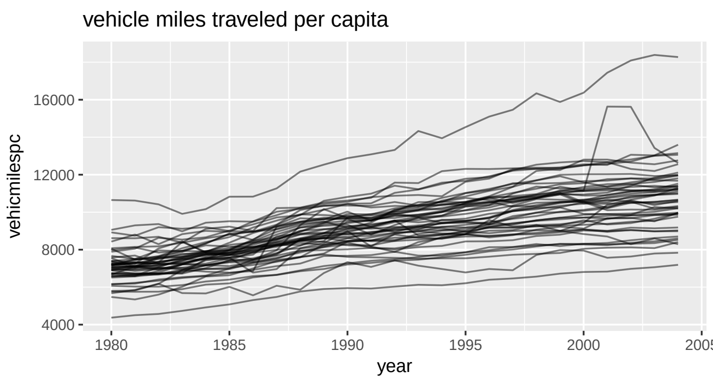
```
## [1] 0
```

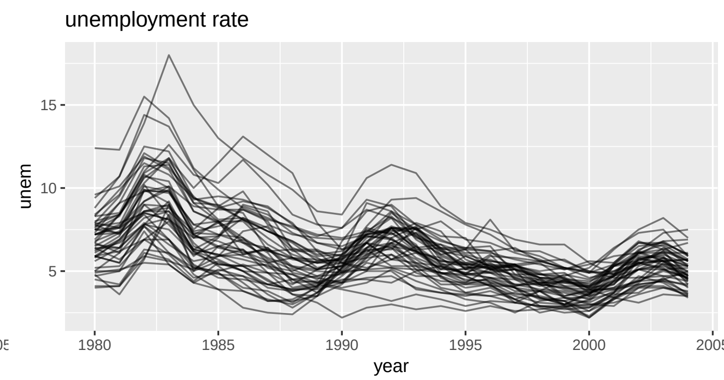
There is no missing data in our data set. Next, we will display the variables we will be using in our model graphically. For the totfartre, perc14_24, vehicmilespc and unem variable, we plot a line for every state over the course of all the years and for the seatbelt, gdl, zerotol, perse, bacXX and slXX variables we create scatter plots (with jitter) that display the distribution accross the different states for every year. To better display the data, we created a combined blood alcohol limit and speed limit variable.

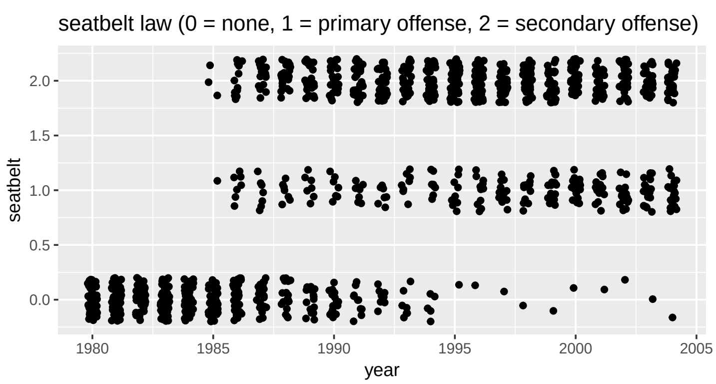
```
plot1 = ggplot(data = data, aes(x = year, y = totfatrte, group = state)) +
  geom_line(alpha=0.5) + ggtitle("total fatality rate")
plot2 = ggplot(data = data, aes(x = year, y = perc14_24, group = state)) +
  geom_line(alpha=0.5) + ggtitle("percentage of population between 14 - 24")
plot3 = ggplot(data = data, aes(x = year, y = vehicmilespc, group = state)) +
  geom_line(alpha=0.5) + ggtitle("vehicle miles traveled per capita")
plot4 = ggplot(data = data, aes(x = year, y = unem, group = state)) +
  geom_line(alpha=0.5) + ggtitle("unemployment rate")
plot5 = ggplot(data = data, aes(x = year, y = seatbelt, group = state)) +
  ggtitle("seatbelt law (0 = none, 1 = primary offense, 2 = secondary offense)") +
  geom_jitter(width = 0.2, height = 0.2)
plot6 = ggplot(data = data, aes(x = year, y = gdl, group = state)) +
  ggtitle("graduate drivers license law") + geom_jitter(width = 0.25, height = 0.1)
plot7 = ggplot(data = data, aes(x = year, y = zerotol, group = state)) +
  ggtitle("zero tolerance law") + geom_jitter(width = 0.25, height = 0.1)
plot8 = ggplot(data = data, aes(x = year, y = perse, group = state)) +
  ggtitle("administrative license revocation (per se law)") +
  geom_jitter(width = 0.25, height = 0.1)
# Create blood alcohol limit variable for law that was valid for the majority of the year.
data$baccombined = ifelse(round(data$bac10) > 0, 0.1, 0.08 * round(data$bac08))
plot9 = ggplot(data = data, aes(x = year, y = baccombined, group = state)) +
  ggtitle("blood alcohol limit for majority of each year") +
  geom_jitter(width = 0.25, height = 0.005)
# Create speed limit variable.
data$slcombined = ifelse(round(data$sl55) > 0, 55, ifelse(round(data$sl65) > 0, 65,
                         ifelse(round(data$sl70) > 0, 70, ifelse(round(data$sl75) > 0, 75, 99))))
plot10 = ggplot(data = data, aes(x = year, y = slcombined, group = state)) +
  ggtitle("speed limit for majority of each year") + geom_jitter(width = 0.25, height = 1)
grid.arrange(plot1, plot2, plot3, plot4, plot5, plot6,
             plot7, plot8, plot9, plot10, ncol=2)
```

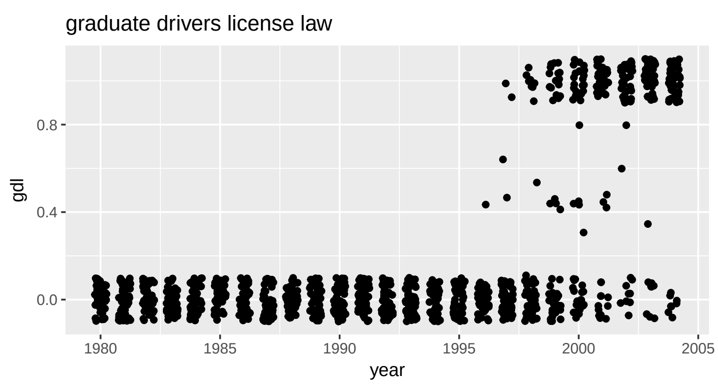

total fatality rate
totfatrte
year

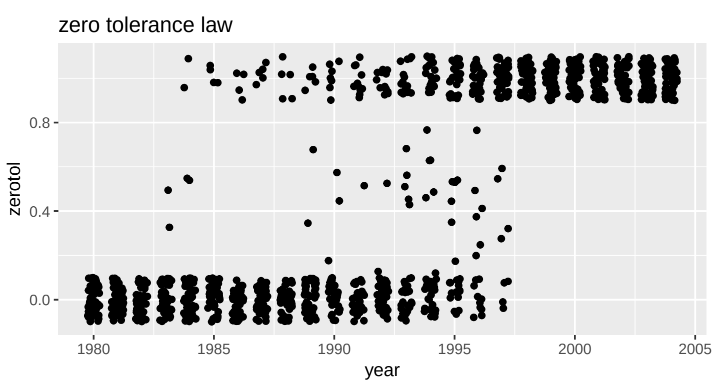

percentage of population between 14 − 24
perc14_24
year

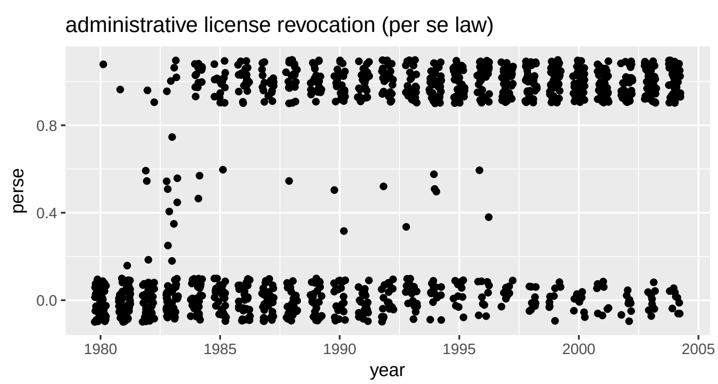

vehicle miles traveled per capita
vehicmilespc
year

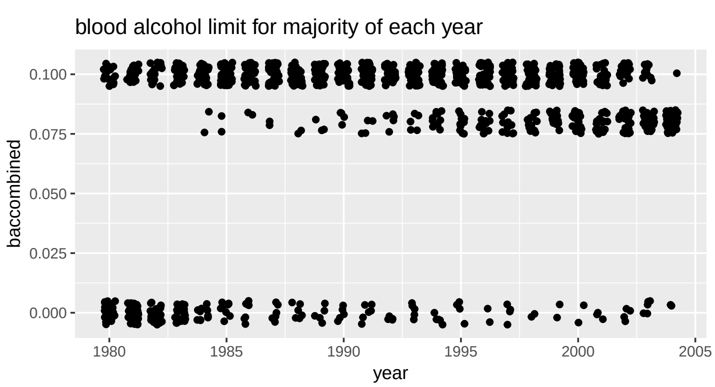

unemployment rate
unem
year

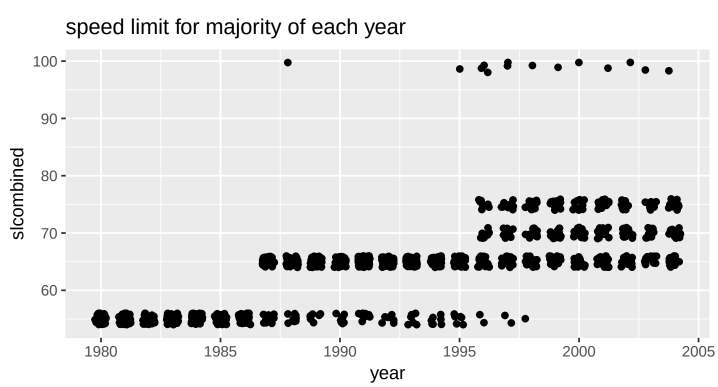

seatbelt law (0 = none, 1 = primary offense, 2 = secondary offense)
seatbelt
year


graduate drivers license law
gdl
year


zero tolerance law
zerotol
year


administrative license revocation (per se law)
perse
year


blood alcohol limit for majority of each year
baccombined
year


speed limit for majority of each year
slcombined
year

For every state, the total fatality rate sank between 1980 and about 1995. Especially in the early eighties there seems to be a steep fall off in the fatality rate. It looks like after 1995 the total fatality rate remained roughly the same for every state.

For 47 out of the 48 continental states under study, the percentage of population between 14 and 24 sank rapidly (from roughly 18% to roughly 14%) between 1980 and and 1990. After that the percentage remained roughly the same.

The vehicle miles traveled per capita increased continuously over the years from 1980 to 2004. There is a strong positive linear trend for each of the states.

The unemployment rate briefly went up in the early eighties and then immediately decreased until the late eighties. In the early nineties the unemployment rate once again increased briefly (on a lower level than before), sank again until around 2000, where it reached its lowest point, before it slightly increase again in the early 00s. It looks like it started decreasing again around 2003.

The first seatbelt laws were implemented in 1985 and by 1995, all but one state had either a primary or secondary offence seatbelt law in place. The distribution remained the same until 2004 (still one state had not implemented a seatbelt law).

Up until the mid 90s, no state had a graduate drivers licence law in place. States began implementing the law in the mid 90s and by 2004, about 80% of all states had implemented a graduate drivers licence law.

In the early 80s, no state had a zero tolerance law in place. The first states that implemented a zero tolerance law did so in the mid 80s. The number of states with zero tolerance laws gradually increased and by the 1998, all states had a zero tolerance law in place.

The minimum drinking age was between 18 and 21 until the mid 80s, when all states implemented a minimum drinking age of 21.

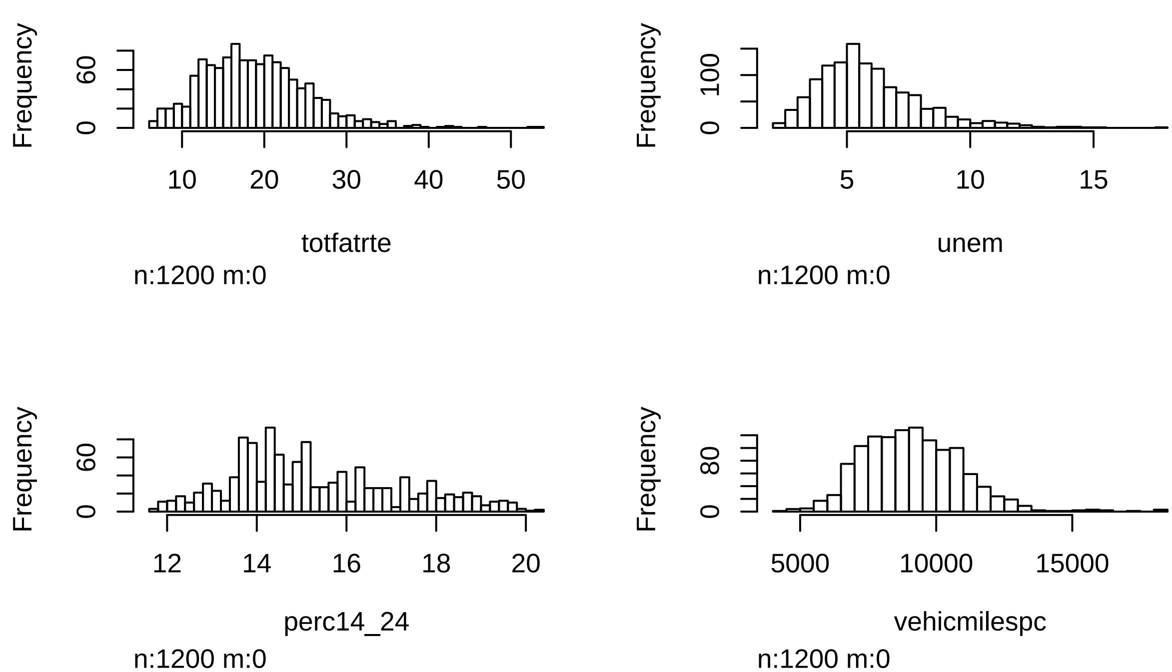
Per se laws were introduced in the early 80 and the amount of states that had per se laws in place increased gradually. By 2004, almost all states had per se laws implemented.

Blood alcohol limit was either 0 or 0.1 in the early 80s. In the mid 80s, they were reduced to 0.08 and more states that had zero as limit set it to 0.08. In 2004 almost all states had a blood alcohol limit of 0.08.

For all states, the speed limits were increased with time. In 1980, all states had a speed limit of 55, wherease by 2004 all states had a speed limit of at least 65, with a substantial number of states having speed limits above.

Next, we take a look at histograms for the non-dummy and non-'fraction of the year' variables that we will be using in our models.

```
hist(data[c('totfatrte','unem','perc14_24','vehicmilespc')])
```



All of these variables are at least a little bit skewed and we will be discussing possible transformation of those in question 3.

ToDo: Remove??? Next, let us take a look at linear correlations between all the non-dummy variables we will be using in our model. Makes no sense for longitudinal data - right?

```
cor(data[c('totfatrte','unem','perc14_24','vehicmilespc','bac08','bac10','perse','sbprim','sbsecon','gdl')])
```

```
##               totfatrte        unem  perc14_24 vehicmilespc       bac08
## totfatrte     1.0000000  0.20670084  0.4017400   0.26654628 -0.16063725
## unem          0.2067008  1.00000000  0.4160272  -0.45389793 -0.20523231
## perc14_24     0.4017400  0.41602716  1.0000000  -0.40016754 -0.27627220
## vehicmilespc  0.2665463 -0.45389793 -0.4001675   1.00000000  0.36089468
## bac08        -0.1606372 -0.20523231 -0.2762722   0.36089468  1.00000000
## bac10        -0.0486040 -0.04360861 -0.1109323  -0.06756487 -0.66374540
## perse        -0.1228792 -0.26757542 -0.4046743   0.44968579  0.40791850
## sbprim       -0.1754423 -0.11620976 -0.2514760   0.13056160  0.24377629
## sbsecon      -0.1068486 -0.32175453 -0.4303547   0.37117624  0.07641566
## gdl          -0.2122301 -0.25468092 -0.2987893   0.31785864  0.44179517
##                   bac10       perse     sbprim     sbsecon        gdl
## totfatrte    -0.04860400 -0.12287920 -0.1754423 -0.10684859 -0.2122301
## unem         -0.04360861 -0.26757542 -0.1162098 -0.32175453 -0.2546809
## perc14_24    -0.11093231 -0.40467427 -0.2514760 -0.43035468 -0.2987893
## vehicmilespc -0.06756487  0.44968579  0.1305616  0.37117624  0.3178586
## bac08        -0.66374540  0.40791850  0.2437763  0.07641566  0.4417952
## bac10         1.00000000 -0.06246905 -0.0466037  0.12092403 -0.2380939
## perse        -0.06246905  1.00000000  0.2155860  0.23402904  0.2744348
## sbprim       -0.04660370  0.21558605  1.0000000 -0.43848923  0.2433590
## sbsecon       0.12092403  0.23402904 -0.4384892  1.00000000  0.1104398
## gdl          -0.23809391  0.27443483  0.2433590  0.11043976  1.0000000
```

# Question 2:

The totfarte variable holds the total fatalities per 100,000 population.

```
df_mean_totfarte = data.frame(aggregate(data$totfatrte, by = list(data$year), mean))
colnames(df_mean_totfarte) <- c("year", "mean_totfarte")
df_mean_totfarte
```

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

```
## 22 2001      16.79271
## 23 2002      17.02958
## 24 2003      16.76354
## 25 2004      16.72896
```

We specify 1980 as the base year and fit the following linear regression model:

$$
\begin{aligned}
totfatrte = &\beta_0 + \beta_1 \cdot d_{81} + \beta_2 \cdot d_{82} + \beta_3 \cdot d_{83} + \beta_4 \cdot d_{84} + \beta_5 \cdot d_{85} + \beta_6 \cdot d_{86} + \beta_7 \cdot d_{87} + \beta_8 \cdot d_{88} + \beta_9 \cdot d_{89} \\
&+ \beta_{10} \cdot d_{90} + \beta_{11} \cdot d_{91} + \beta_{12} \cdot d_{92} + \beta_{13} \cdot d_{93} + \beta_{14} \cdot d_{94} + \beta_{15} \cdot d_{95} + \beta_{16} \cdot d_{96} + \beta_{17} \cdot d_{97} \\
&+ \beta_{18} \cdot d_{98} + \beta_{19} \cdot d_{99} + \beta_{20} \cdot d_{00} + \beta_{21} \cdot d_{01} + \beta_{22} \cdot d_{02} + \beta_{23} \cdot d_{03} + \beta_{24} \cdot d_{04}
\end{aligned}
$$

```r
lm1 = 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(lm1)
```

```
##
## 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
```

All parameter estimates except the one for 1981 are highly statistically significant.

This model explains exactly what was calculated in the aggregation above. If we add any of the above estimated dummy parameters to the intercept, we receive the mean totfatrte value for the respective year. Below, we show this for three randomly selected years.

```
set.seed(999)
results = data.frame()
for (i in 1:25) {
  mean_value_lm1 = ifelse(i==1, lm1$coefficients[1], lm1$coefficients[1] + lm1$coefficients[i])
  year = df_mean_totfarte$year[i]
  mean_value_agg = df_mean_totfarte$mean_totfarte[i]
  difference = round(mean_value_lm1 - mean_value_agg,5)
  results = rbind(results, cbind(year, mean_value_agg, mean_value_lm1, difference))
}
sample_n(results, 3)
```

```
##    year mean_value_agg mean_value_lm1 difference
## 10 1989       19.77229       19.77229          0
## 14 1993       17.12771       17.12771          0
## 3  1982       20.94250       20.94250          0
```

We looked at the residual plot for this model and saw that the residuals are not well behaved. This model is not taking the nested dependency (accross states) within the data into account. We furthermore strongly believe that we are missing a lot of important explanatory variables.

When we take a look at the mean values of total traffic fatalities between 1980 and 2004, we see that there was certainly an improvement over time. The improvement however did not occur 'continuously' (year by year), but rather in two steps. The first improvement took place between 1980 and 1983 (decrease of roughly 5.3) and the second improvement took place between 1990 and 1992 (decrease of around 2.3). During the other time periods, the total traffic fatalaties remained roughly constant every year (with just some 'natural' variation).

## Question 3:

Like in question 2, we are fitting a pooled OLS linear regression model, but this time we are including more explanatory variables than just the dummy variables for the years. We are including bac08 and bac10, perse, sbprim and sbsecond, sl70plus, gdl, perc14_24, unem and vehicmilespc.

Most of our newly added variables are either dummy variables (sbprim, sbsecond) or indicate fractions of a year (bac08, bac10, perse, sl70plus, gdl), which mainly hold either 0 or 1. Naturally, the distributions for these variables all have extreme peaks at 0 and 1.

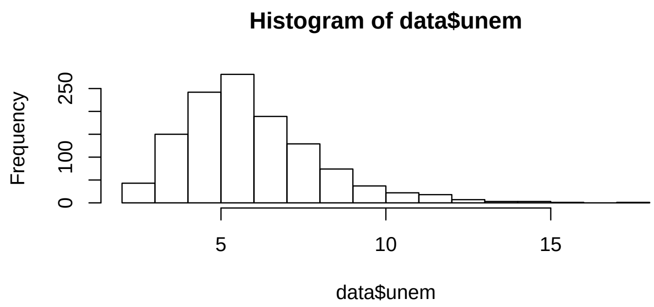
We were debating on whether or not to bin the 'fraction of year' variables to 0 and 1. We however believe that the information loss outweighs the gain in comprehensibility/simplicity, which is why we decided to leave the variables as they are.

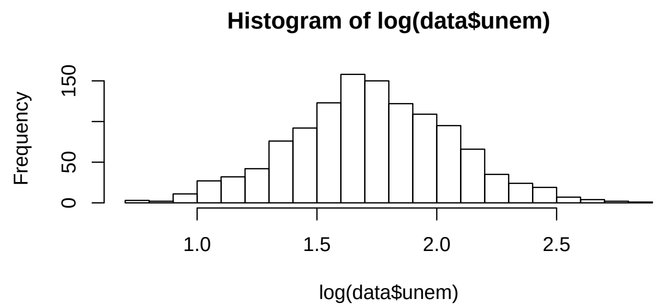
For the other variables (totfatrte, perc14_24, unem and vehicmilespc) we looked at the histograms and based on that tried out different transformations. We furthermore performed Shapiro Wilk Tests on the distributions before and after transformation to check if the transformation improved the distribution, in terms of being closer to a normal distribution. The Shapiro Wilk Test tests against the null hypothesis, that the variable stems from a normal distribution.

The totfatrte variable has a right skew. We were not able to find a simple transformation that significantly improved the p-value of the Shapiro Wilk Test, which is why we decided to not transform this variable. The distribution of the perc14_24 variable has an abnormal peak around 14, but looks well balanced otherwise. Different transormations did not improve the distribution significantly. For every transformation, the Shapiro Wilk Test for the log transformed variable continued to show high statistical significance. The distribution of the unem variable has a strong right skew which can be beautifully removed using a log transformation. The resulting log(unem) variable closely resembles a normal distribution and this is also confirmed by the Shapiro Wilk Test. The p-value is close to 0.3 and hence we fail to reject the null hypothesis, which means that it is likely that log(unem) stems from a normal distribution (see analysis below). The distribution of the vehicmilespc variable is quite a bit leptokurtic, but balanced otherwise. While a log transformation significantly increased the p-value in the Shapiro Wilk Test, the p-value is still far away from being not significant. We therefore decided to not perform a transformation on the vehicmilespc variable.

```
par(mfrow = c(1, 2))
hist(data$unem, breaks = 20)
```

```r
hist(log(data$unem), breaks = 20)
```




```r
shapiro.test(data$unem)$p.value
```

```
## [1] 1.265313e-22
```

```r
shapiro.test(log(data$unem))$p.value
```

```
## [1] 0.2991468
```

We are therefore specifying the following model:

$$
\begin{aligned}
totfatrte = &\beta_0 + \beta_1 \cdot d_{81} + \beta_2 \cdot d_{82} + \beta_3 \cdot d_{83} + \beta_4 \cdot d_{84} + \beta_5 \cdot d_{85} + \beta_6 \cdot d_{86} + \beta_7 \cdot d_{87} + \beta_8 \cdot d_{88} + \beta_9 \cdot d_{89} \\
&+ \beta_{10} \cdot d_{90} + \beta_{11} \cdot d_{91} + \beta_{12} \cdot d_{92} + \beta_{13} \cdot d_{93} + \beta_{14} \cdot d_{94} + \beta_{15} \cdot d_{95} + \beta_{16} \cdot d_{96} + \beta_{17} \cdot d_{97} \\
&+ \beta_{18} \cdot d_{98} + \beta_{19} \cdot d_{99} + \beta_{20} \cdot d_{00} + \beta_{21} \cdot d_{01} + \beta_{22} \cdot d_{02} + \beta_{23} \cdot d_{03} + \beta_{24} \cdot d_{04} + \beta_{25} \cdot bac08 \\
&+ \beta_{26} \cdot bac10 + \beta_{27} \cdot perse + \beta_{28} \cdot sbprim + \beta_{29} \cdot sbsecon + \beta_{30} \cdot gdl + \beta_{31} \cdot perc_{14\_24} + \\
&+ \beta_{32} \cdot \log(unem) + \beta_{33} \cdot vehicmilespc
\end{aligned}
$$

```r
lm2 = 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 +
         bac08 + bac10 + perse + sbprim + sbsecon + gdl + perc14_24 + log(unem) +
         vehicmilespc, data = data)
summary(lm2)
```

```
##
## Call:
## lm(formula = totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 +
##     d87 + d88 + d89 + d90 + d91 + d92 + d93 + d94 + d95 + d96 +
##     d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04 + bac08 + bac10 +
##     perse + sbprim + sbsecon + gdl + perc14_24 + log(unem) +
##     vehicmilespc, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.9019  -2.6583  -0.4533   2.3591  21.5414
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.453e+01  2.507e+00  -5.795 8.76e-09 ***
## d81           -2.039e+00  8.394e-01  -2.429 0.015289 *
## d82           -6.202e+00  8.566e-01  -7.240 8.12e-13 ***
## d83           -6.876e+00  8.717e-01  -7.888 7.03e-15 ***
## d84           -5.399e+00  8.878e-01  -6.081 1.61e-09 ***
## d85           -5.962e+00  9.063e-01  -6.579 7.14e-11 ***
## d86           -5.132e+00  9.431e-01  -5.441 6.45e-08 ***
## d87           -5.471e+00  9.811e-01  -5.576 3.05e-08 ***
## d88           -5.471e+00  1.030e+00  -5.309 1.32e-07 ***
## d89           -6.908e+00  1.069e+00  -6.463 1.51e-10 ***
```

```
## d90          -7.871e+00  1.092e+00  -7.211 9.99e-13 ***
## d91          -1.004e+01  1.115e+00  -9.010  < 2e-16 ***
## d92          -1.185e+01  1.136e+00 -10.436  < 2e-16 ***
## d93          -1.166e+01  1.151e+00 -10.130  < 2e-16 ***
## d94          -1.117e+01  1.173e+00  -9.516  < 2e-16 ***
## d95          -1.061e+01  1.202e+00  -8.823  < 2e-16 ***
## d96          -1.124e+01  1.216e+00  -9.240  < 2e-16 ***
## d97          -1.094e+01  1.229e+00  -8.897  < 2e-16 ***
## d98          -1.143e+01  1.243e+00  -9.199  < 2e-16 ***
## d99          -1.136e+01  1.265e+00  -8.978  < 2e-16 ***
## d00          -1.158e+01  1.290e+00  -8.982  < 2e-16 ***
## d01          -1.277e+01  1.313e+00  -9.725  < 2e-16 ***
## d02          -1.362e+01  1.325e+00 -10.281  < 2e-16 ***
## d03          -1.397e+01  1.337e+00 -10.454  < 2e-16 ***
## d04          -1.347e+01  1.367e+00  -9.854  < 2e-16 ***
## bac08        -2.626e+00  5.454e-01  -4.815 1.67e-06 ***
## bac10        -1.402e+00  4.020e-01  -3.487 0.000506 ***
## perse        -4.423e-01  3.019e-01  -1.465 0.143122
## sbprim       -2.954e-01  5.002e-01  -0.591 0.554853
## sbsecon       3.013e-02  4.359e-01   0.069 0.944917
## gdl          -3.925e-01  5.348e-01  -0.734 0.463097
## perc14_24     4.346e-01  1.194e-01   3.641 0.000283 ***
## log(unem)     5.526e+00  4.875e-01  11.336  < 2e-16 ***
## vehicmilespc  3.061e-03  9.404e-05  32.551  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.105 on 1166 degrees of freedom
## Multiple R-squared:  0.5958, Adjusted R-squared:  0.5844
## F-statistic: 52.08 on 33 and 1166 DF,  p-value: < 2.2e-16
```

The estimated model parameter is roughly -2.6 for the bac08 variable and roughly -1.4 for bac10 variable. This means the model suggests that (ceteris paribus), when a blood alcohol limit of 0.08 is introduced, states can expect to lower their total traffic fatalities rate per 100,000 by about 2.6 per year. If (ceteris paribus) a blood alcohol limit of 0.1 is introduced, states can expect to lower their total traffic fatalities rate per 100,000 by about 1.4 per year. This result is surprising and certainly counter intuitive. By looking at our graphical representation of the combined bac08 and bac10 variable in the EDA section, this result can however be explained: While only a fraction of the states introduced a blood alcohol limit of 0.1 throughout the period of our study, it looks like (almost) all states implemented a 0.08 blood alcohol limit in 2004 (possibly a federal law) - and quite a few states even did so before. This means: - When states implemented a 0.1 limit, they always RAISED the blood alcohol limit (from either 0 or 0.08 to 0.1) and we would expect a decrease in the traffic fatalities. - However, when states implemented a 0.08 limit, they either RAISED the blood alcohol limit from 0 to 0.08 (we would epect a decrease in the traffic fatalities) or they LOWERED it from a previously existing 0.1 blood alcohol limit (we would expect an increase in the traffic fatalities). As the model doesn't account for the historic state in terms of blood alcohol limit, the parameters are 'skewed'. That is especially true for the 0.08 blood alcohol limit, as the change in limit can be in either direction, whereas for the 0.1 limit it always went in the same direction (and is only of different magnitude). In this instance, pooled OLS is a poor choice of model, as it fails to adequately capture the time-varying effects of either `bac` variable.

According to our model, an implementation of per se laws slightly lowers the fatality rate. For a full-year implementation of a per se law we would expect a decrease in the total fatalities rate of about 0.44. This result however is not statistically significant and therefore should not be taken at face value.

According to our model, the introduction of a primary seat belt law lowers the fatality rate. For a full-year implementation of a primary seat belt law, we would expect a decrease in the total fatalities rate of about 0.3. Again, we have the problem that certain states might have switched from a secondary to a primary seatbelt law, which would skew this parameter. The result however is far away from statistical significane and therefore should not be taken at face value.

# Question 4:

We reestimate the model from Question 3 using a Fixed Effects Model at the state level. By specifiying a Fixed Effects Model, we account for the time invariant unobserved hetereogenity.

```r
fe1 = plm(totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89 + d90 + d91 +
          d92 + d93 + d94 + d95 + d96 + d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04 +
          bac08 + bac10 + perse + sbprim + sbsecon + gdl + perc14_24 + log(unem) +
          vehicmilespc, data = data, model = 'within', index = c('state'))
summary(fe1)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 +
##     d87 + d88 + d89 + d90 + d91 + d92 + d93 + d94 + d95 + d96 +
##     d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04 + bac08 + bac10 +
##     perse + sbprim + sbsecon + gdl + perc14_24 + log(unem) +
##     vehicmilespc, data = data, model = "within", index = c("state"))
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -8.357812 -1.030083 -0.025029  0.953108 14.607455
##
## Coefficients:
##                 Estimate  Std. Error  t-value  Pr(>|t|)
## d81          -1.5767e+00  4.1295e-01  -3.8182 0.0001418 ***
## d82          -3.3444e+00  4.3290e-01  -7.7256 2.464e-14 ***
## d83          -3.8734e+00  4.4583e-01  -8.6881 < 2.2e-16 ***
## d84          -4.4190e+00  4.6261e-01  -9.5523 < 2.2e-16 ***
## d85          -4.8760e+00  4.8218e-01 -10.1124 < 2.2e-16 ***
## d86          -3.8891e+00  5.1355e-01  -7.5729 7.616e-14 ***
## d87          -4.5718e+00  5.5166e-01  -8.2873 3.285e-16 ***
## d88          -5.0927e+00  5.9947e-01  -8.4954 < 2.2e-16 ***
## d89          -6.4173e+00  6.3610e-01 -10.0885 < 2.2e-16 ***
## d90          -6.4571e+00  6.5801e-01  -9.8131 < 2.2e-16 ***
## d91          -7.1320e+00  6.7207e-01 -10.6119 < 2.2e-16 ***
## d92          -8.0188e+00  6.9234e-01 -11.5823 < 2.2e-16 ***
## d93          -8.3327e+00  7.0629e-01 -11.7978 < 2.2e-16 ***
## d94          -8.7699e+00  7.2634e-01 -12.0741 < 2.2e-16 ***
## d95          -8.5791e+00  7.5032e-01 -11.4340 < 2.2e-16 ***
## d96          -8.9618e+00  7.6604e-01 -11.6988 < 2.2e-16 ***
## d97          -9.1765e+00  7.8153e-01 -11.7418 < 2.2e-16 ***
## d98          -9.9090e+00  7.9589e-01 -12.4502 < 2.2e-16 ***
## d99          -1.0117e+01  8.1004e-01 -12.4893 < 2.2e-16 ***
## d00          -1.0712e+01  8.2569e-01 -12.9736 < 2.2e-16 ***
## d01          -1.0127e+01  8.3511e-01 -12.1270 < 2.2e-16 ***
## d02          -9.2601e+00  8.4154e-01 -11.0037 < 2.2e-16 ***
## d03          -9.2604e+00  8.4944e-01 -10.9017 < 2.2e-16 ***
## d04          -9.7507e+00  8.7443e-01 -11.1509 < 2.2e-16 ***
## bac08        -1.3185e+00  3.9535e-01  -3.3351 0.0008808 ***
## bac10        -9.6089e-01  2.6936e-01  -3.5673 0.0003759 ***
## perse        -1.2151e+00  2.3232e-01  -5.2300 2.022e-07 ***
## sbprim       -1.1696e+00  3.4259e-01  -3.4140 0.0006632 ***
## sbsecon      -2.9761e-01  2.5208e-01  -1.1806 0.2380021
## gdl          -3.8805e-01  2.9257e-01  -1.3264 0.1849901
## perc14_24     1.6393e-01  8.9621e-02   1.8291 0.0676448 .
## log(unem)    -3.6604e+00  3.9188e-01  -9.3406 < 2.2e-16 ***
## vehicmilespc  9.6380e-04  1.1003e-04   8.7594 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Total Sum of Squares:    12134
## Residual Sum of Squares: 4543.7
## R-Squared:      0.62554
## Adj. R-Squared: 0.59877
## F-statistic: 56.646 on 33 and 1119 DF, p-value: < 2.22e-16
```

ToDo: How do the coefficients on *bac08, bac10, perse, and sbprim* compare with the pooled OLS estimates? Which set of estimates do you think is more reliable? What assumptions are needed in each of these models? Are these assumptions reasonable in the current context?

lm model output of the four variables under study.

bac08 -2.626e+00 5.454e-01 -4.815 1.67e-06 ***bac10 -1.402e+00 4.020e-01 -3.487 0.000506*** perse -4.423e-01 3.019e-01 -1.465 0.143122
sbprim -2.954e-01 5.002e-01 -0.591 0.554853

fe model output of the four variables under study.

bac08 -1.3185e+00 3.9535e-01 -3.3351 0.0008808 ***bac10 -9.6089e-01 2.6936e-01 -3.5673 0.0003759*** perse -1.2151e+00 2.3232e-01 -5.2300 2.022e-07 ***sbprim -1.1696e+00 3.4259e-01 -3.4140 0.0006632***

## Question 5:

To determine wheter or not a random effects model should be used instead of the fixed effects model, we first conduct a Hausman Test, with null hypothesis that the random effects assumptions are correct.

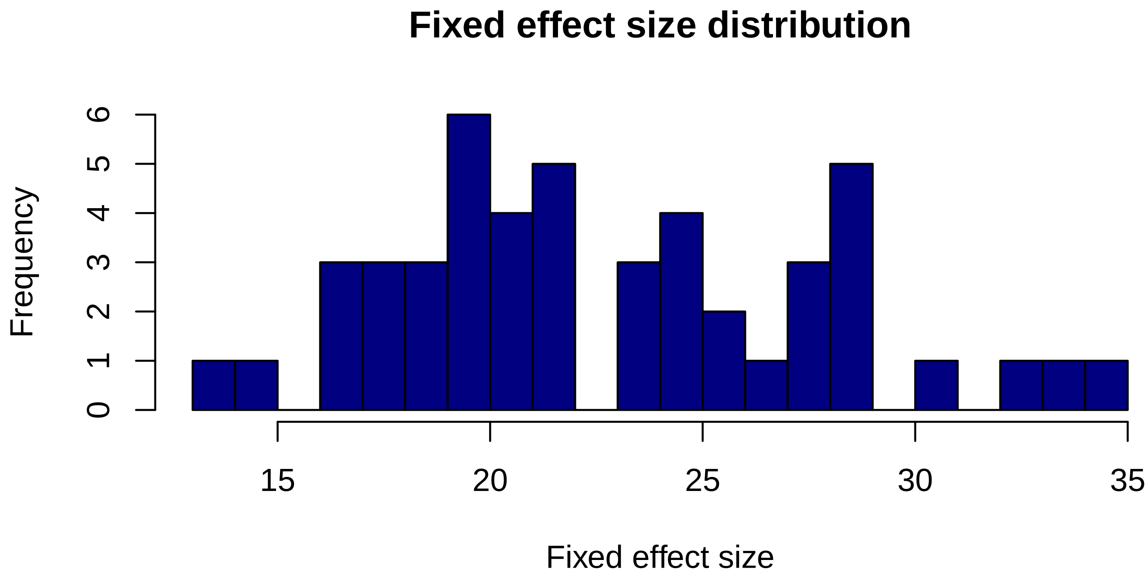
```r
re1 = plm(totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89 + d90 + d91 +
          d92 + d93 + d94 + d95 + d96 + d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04 +
          bac08 + bac10 + perse + sbprim + sbsecon + gdl + perc14_24 + log(unem) +
          vehicmilespc, data = data, model = 'random', index = c('state'))
phtest(fe1, re1)
```

```
##
##  Hausman Test
##
## data:  totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89 +  ...
## chisq = 78.454, df = 33, p-value = 1.436e-05
## alternative hypothesis: one model is inconsistent
```

We reject the null hypothesis with a highly statistically significant p-value. The Hausman Test therefore suggests to use a Fixed Effects model.

Since furthermore none of our explanatory variables are constant over time (which can't be modeled using fixed effects) and we observe quite a bit of variability in the estimated fixed effects for our `fe1` model (see histogram of the fixed effect accross the states below), we believe that the Fixed Effects Model is the better model for our scenario.

```r
#Fixed effects of fe1 model
hist(fixef(fe1), breaks = 20, c='navy', xlab='Fixed effect size',
     main='Fixed effect size distribution')
```




## Question 6:

According to our Fixed Effects model we can state the following:

If - on average - people were to drive 1,000 miles more per year, we would expect the rate of total traffic fatalaties per 100,000 people to increase by roughly 1 (0.96 according to model).

Let us explain this with an example. In 2004, the United States had a popluation of roughly 300 million people. If - on average - people had driven 1,000 miles more in the year of 2004 than they actually did, we would expect about 2,880 additional traffic fatalities.

## Question 7:

If there is serial correlation in the idiosyncratic errors of the model, testable using a function such as `pbgtest` from the `plm` package, the estimated coefficients will be biased [https://ageconsearch.umn.edu/bitstream/116069/2/sjart_st0039.pdf], meaning the relationship between the predictor and outcome variables will be mischaracterized. Further, the estimated variance will be biased, resulting in an invalid estimate for the standard error as well.

If there is heteroskedasticity in the idiosyncratic errors, the coefficients may not necessarily be biased, but this condition will cause the estimated variance to become biased, resulting in an invalid estimate for the standard error [https://en.wikipedia.org/wiki/Heteroscedasticity#Consequences].