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

W271 Instructional Team

Fall 2018

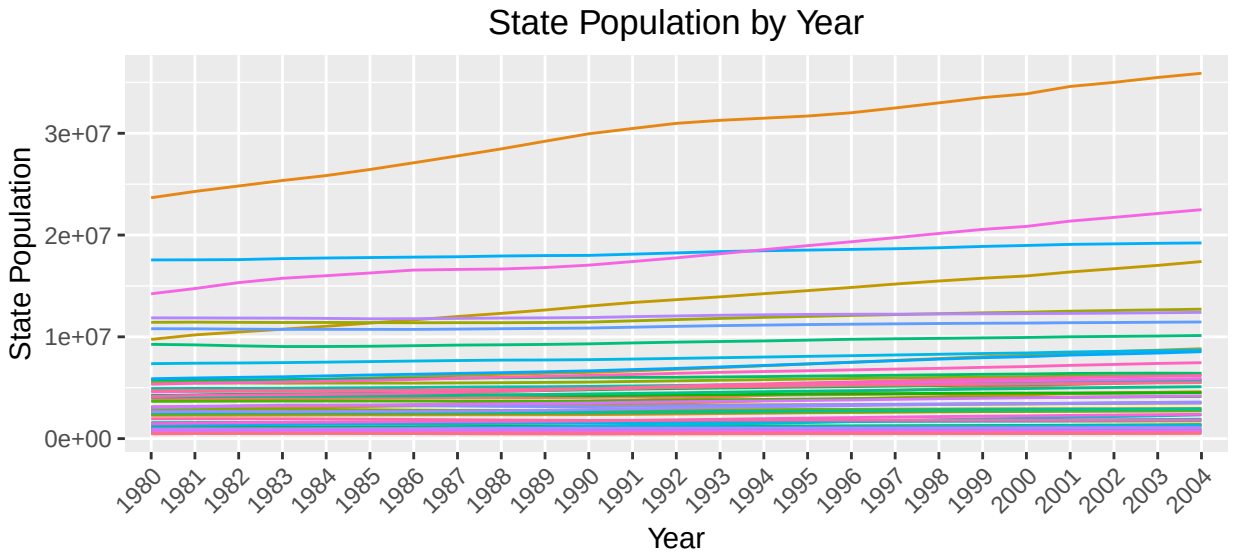
```
load("driving.RData")
```

(Questsion 1) The dataset contains about 1200 observations ranging from 1980 to 2004 for the 48 continental states. The observed variables include: Speed limits (`slXX`) seat belt and zero tolerance, graduated driver, blood alcohol level (`bacXX`), per se are in percent of year by months in binary. `Sbl70plus`, `sbprim`, `sbsecon` and `dXX` variables are simply derivatives or dummy variables of the other variables in the data set. Other variables are continuous with a a base of 0 and no top coding, except perc14_24 (100%).

Our research question is whether or not traffic laws can affect total fatalities. Total fatalities is a function of population, vehicle miles, traffic laws and unobservable variables. Our dataset contains 9 fatality-related variables; some normalized in various ways. We will not consider the weekend and night fatality variables as we are focused on total fatalities and not when they occurred.

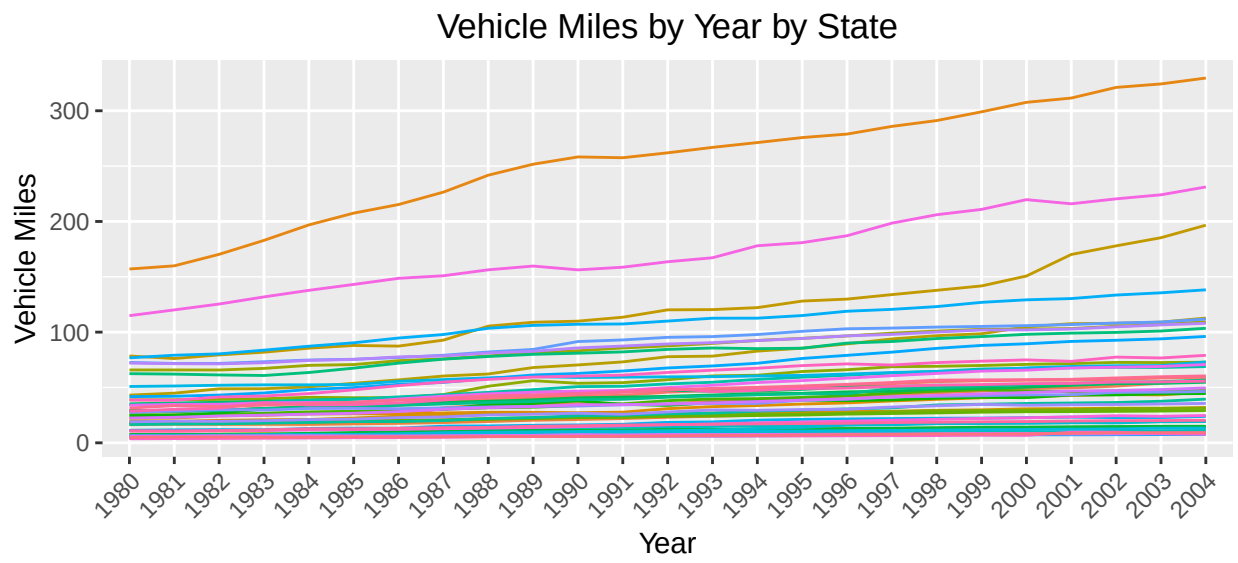
(Question 2) Three states, 5, 44 and 10, significantly increased in population while other states were relatively flat. This suggests using a population normalized fatality measure, such as totfatrte, for examining traffic laws

```
ggplot(data, aes(y = statepop, x = factor(year), group = factor(state), color = factor(state))) +
    geom_line() + theme(axis.text.x = element_text(angle = 45, hjust = 1)) + labs(x = c("Year"),
    y = c("State Population"), title = c("State Population by Year"), color = c("State")) +
    theme(plot.title = element_text(hjust = 0.5)) + guides(color = FALSE)
```



(Question 2) Vehicle miles has roughly similar trends for almost all states. This further confirms totfatrte as a dependent variable for traffic law causal inference as vehicle miles is more "stable" among the states than population through out time.

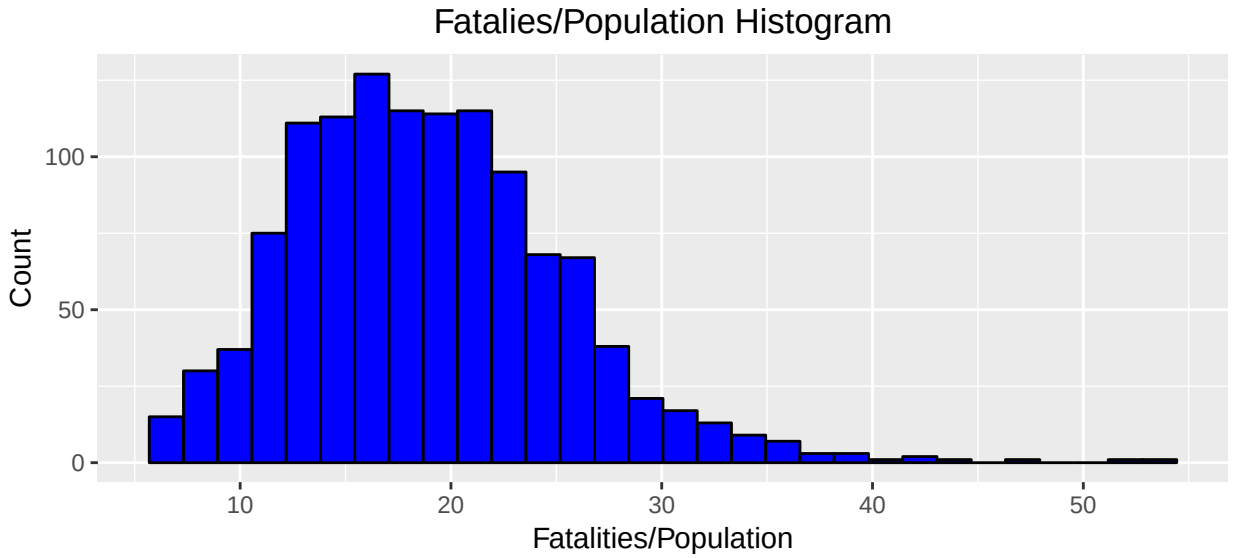
```
ggplot(data, aes(y = vehicmiles, x = factor(year), group = factor(state), color = factor(state))) +
    geom_line() + theme(axis.text.x = element_text(angle = 45, hjust = 1)) + labs(x = c("Year"),
    y = c("Vehicle Miles"), title = c("Vehicle Miles by Year by State"), color = c("State")) +
    theme(plot.title = element_text(hjust = 0.5)) + guides(color = FALSE)
```

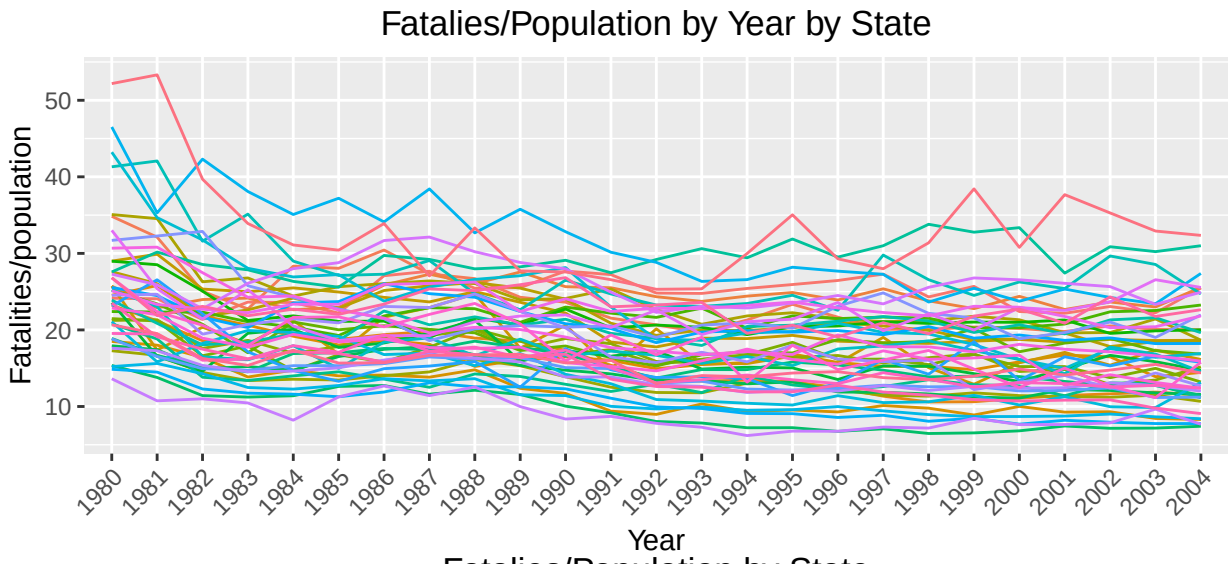


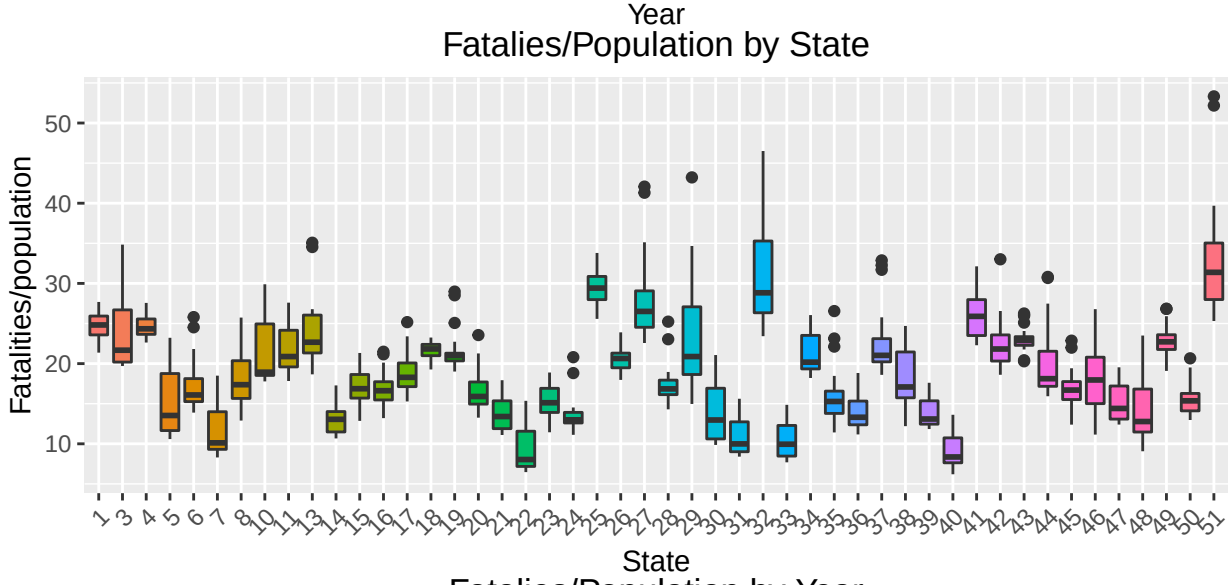
(Question 2) Mean fatalities/population has dropped by about 10% from 1980 to 2004. State 51 has persistently stayed near the top of the fatalies while State 38 seems to stay near the bottom. The range of fatalities changes throughout time, with the highest fatality rate dropping from over 50 in 1980 to under 45 in 2004. The minimum values for fatalities drop at well, but less overall, from about just over 10 in 1980 to just under 10 in 2004.

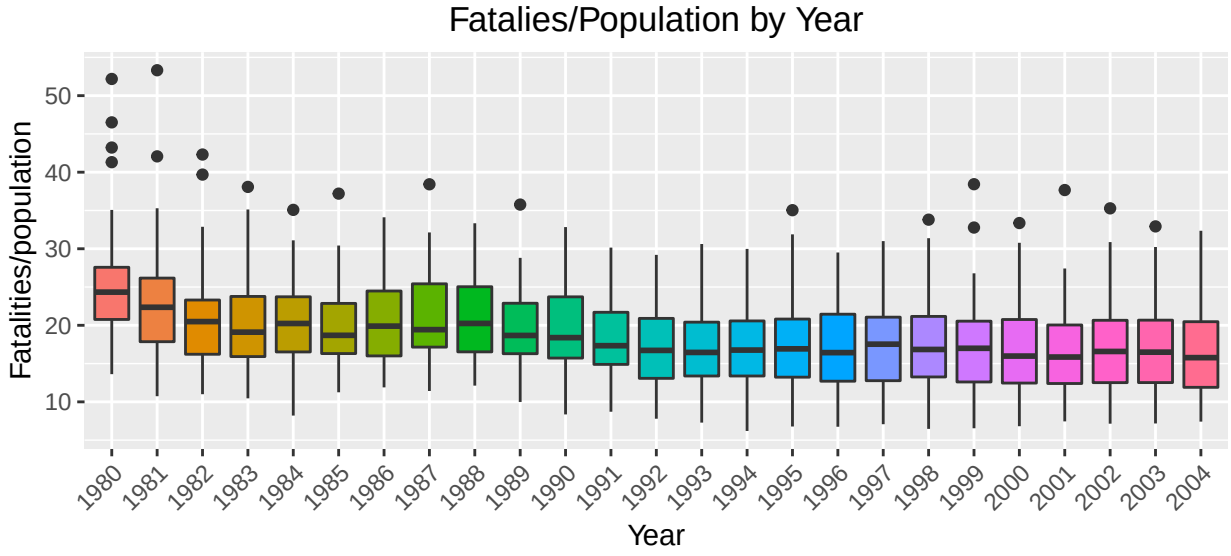
There are several states that have higher overall fatality rates than the others consistently - states 25, 32 and 51 may be further explored to see the relationships with the observed variables.

```
plottot("totfatrte")
```




Fatalies/Population by Year by State
Fatalities/population
50
40
30
20
10
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

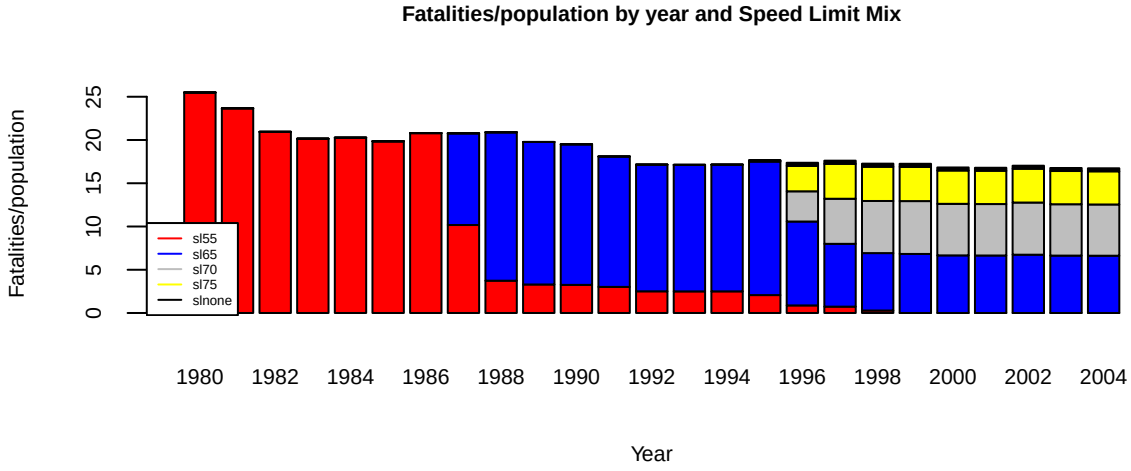

Fatalies/Population by State
Fatalities/population
50
40
30
20
10
1 3 4 5 6 7 8 10 11 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51
State


Fatalies/Population by Year
Fatalities/population
50
40
30
20
10
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

SS I'm a little confused as to who these plot are put together. same question on all stacked-like bar charts. Could we view stacked bar charts of the Speed limits and impose and average line for `totfatrte` on top

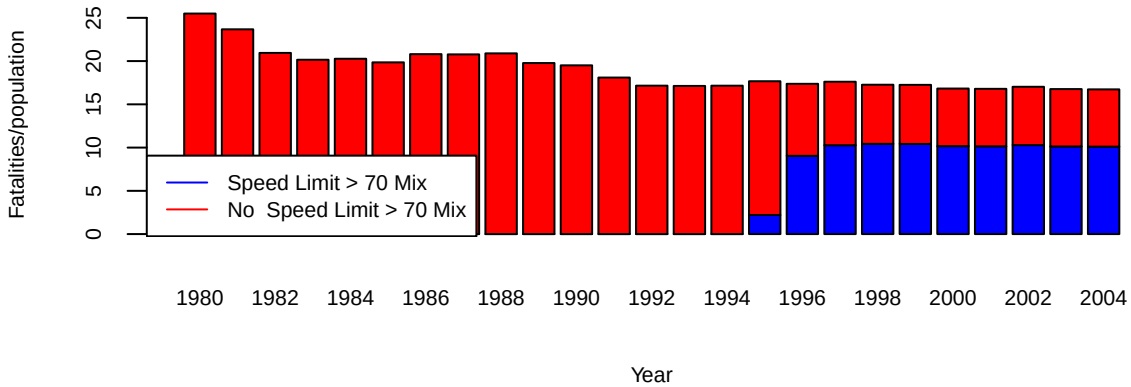
(Question 1) Initially, the speed limit (`slXX`) increase does not seem to affect the average total fatality rate across states immediately. In 1986-1987, speed limits increased in many states. In 1988-1991, fatalities fell by about 10%. The variable may be a candidate as an interaction variable with time.

```
plotstackedbox_sl()
```



To further examine speed limits, we split the variables into two groups - states with speed limits under 70 and states with speed limits over 70. From the chart below, it appears that speed limit over 70+ do not impact fatalities in a significant way.

```
plotstackedbox(data, data$sl70plus, "Speed Limit > 70", "Speed Limit > 70 Mix")
```



`bac_none` represents where states did not have laws relating to blood alcohol content.

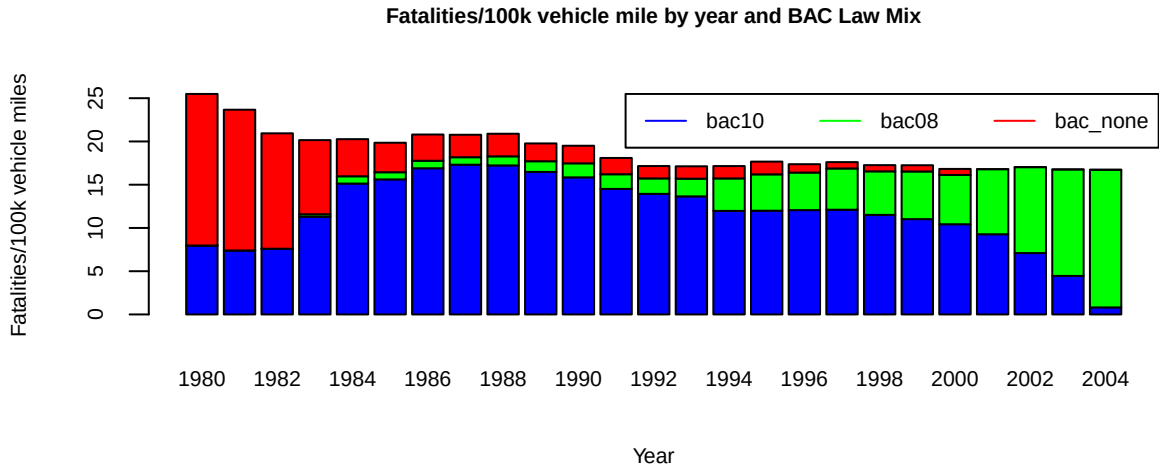
Blood Alcohol Content laws appear to have an immediate impact on fatalities. More interestingly, the graph suggests transformig the BAC laws into a binary variable, as a `bac08` and `bac10` does not appear to affect fatalities, but the initial implementation of drinking-related laws has an impact.

We performed cross-sectional t-tests for 2001 and 2002 when the `bac08` and `bac10` were closest to 50% between the states. The two t-tests are performed to avoid the general downward trend of fatalities impacting the analysis. In both t-tests, the H_0: differences in means = 0 are not rejected. From the graph below, there aren't likely to be lagged effects for `bac10` and `bac08` as `bac10` decreased from 1997-2004, but fatalities

appear roughly the same. The lack of lagged effect makes sense as bac laws will immediately impact drunk driver and remove them from the roads.

We may later test the joint probability of `bac08` and `bac10`. If both show insignificance in a multivariate regression but rejects the H_0 jointly, we should convert it into a binary variable of BAC laws or none.

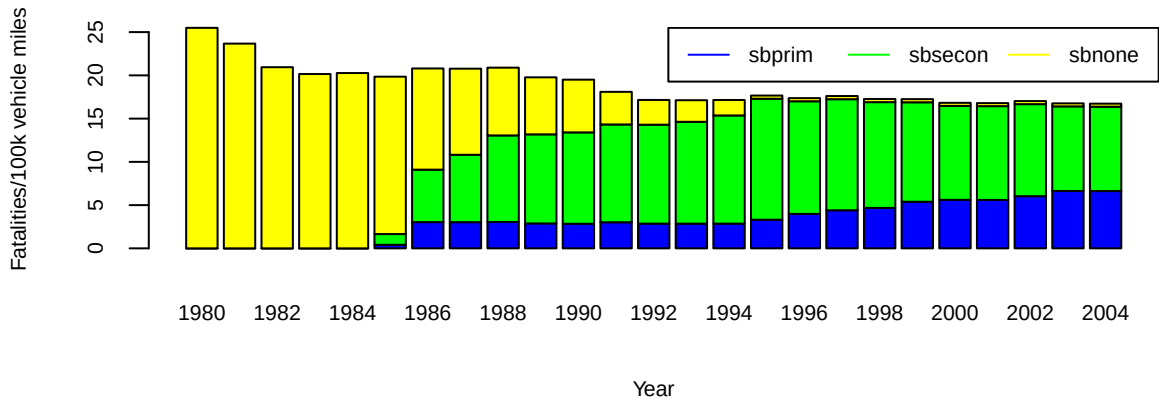
```
plotstackedbox_bac()
```



Seatbelts appear to have a simultanenous decrease with fatalities. It should be included as an independent variable. Much like BAC levels, the mix of seatbelt laws does not appear to affect fatalities. There appears to be some lagged effect on the binary seatbelt law - possibly the population is getting in the habit of putting on seatbelts. After most states have implemented at least primary seatbelt laws (1992-1995), though, the average fatality rate flattens.

We performed cross-sectional t-tests for 1999 and 2000 where the percentages are closer to even for sbprim and sbsecon. For both t-tests, H_0 is not rejected and there does not appear to be any contemporaneous differences between seatbelt laws. Despite the decrease of sbsecon mix from 1995 to 2004, the fatalities from 1995 to 2004 is similar. There is unlikely to be a lagged impact of seatbelt law differences. Finally, we also note that state 30 does not have seatbelt laws throughout the period if further analysis on specifics of the seatbelt law is required.

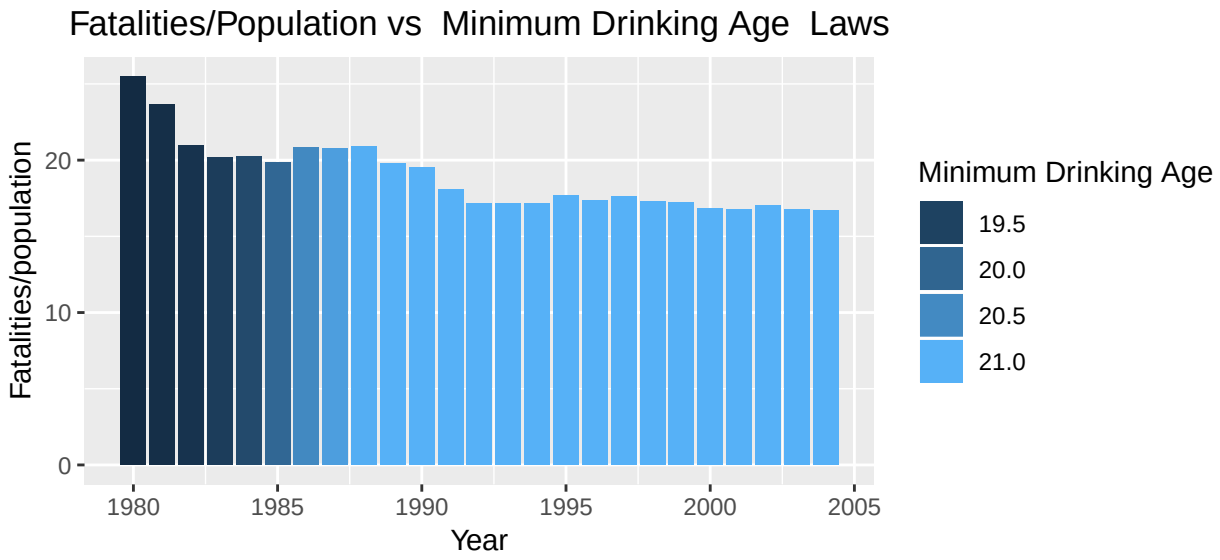
```
plotstackedbox_seatbelt()
```



Minimum drinking age appears to have some effect on fatalies. Interestingly, all states listed did not have minimum age laws changed during the period. The mean `minage` from 1980-1990 trended higher to 21 in 1990. If we were to focus on `minage`, we can split the data into 2 sets and run separate analysis, detrend

and analyze the impact of `minage` on fatalities. `minage` could be interacted this variable with BAC variables as raising minimum drinking age may potentially offset some effect of BAC laws. The interaction term is expected to have a negative coefficient.

```
plotmix(data, "minage", "Minimum Drinking Age")
```

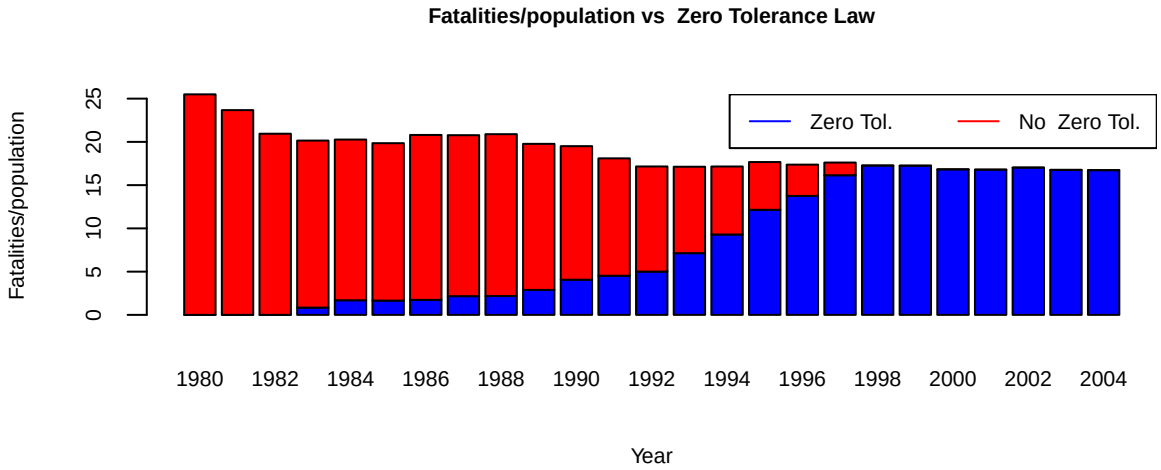



```
tmp2 = data %>% group_by(state) %>% summarize_all(funs(mean)) %>% as.data.frame
tmp2 = tmp2[tmp2$minage %in% unique(data$minage), c("state", "minage")]
tmp2 = tmp2[!(tmp2$state %in% c(47, 51)), c("state", "minage")]
xtable(t(tmp2))
```

	3	4	11	12	15	20	23	26	29	32	35	36	42	45
state	4.00	5.00	14.00	15.00	18.00	23.00	26.00	29.00	32.00	35.00	38.00	39.00	45.00	48.00
minage	21.00	21.00	21.00	21.00	21.00	21.00	21.00	21.00	21.00	21.00	21.00	21.00	21.00	21.00

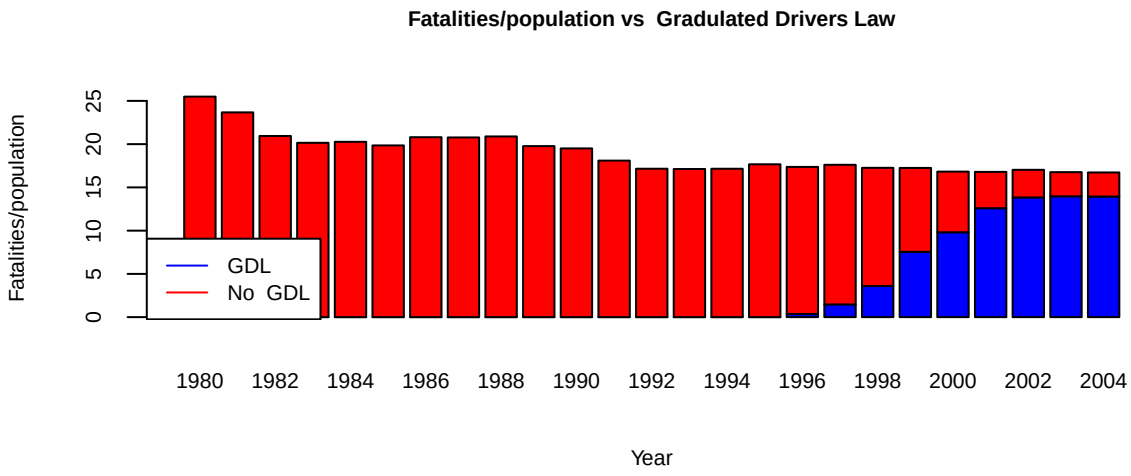
Zero tolerance laws do not appear to have a contemperous impact on fatalities based on the changes in laws from 1992-1997. It may potentally have a long-tailed effect.

```
plotstackedbox(data, data$zerotol, "Zero Tolerance Law", "Zero Tol. ", "topright",
    2)
```



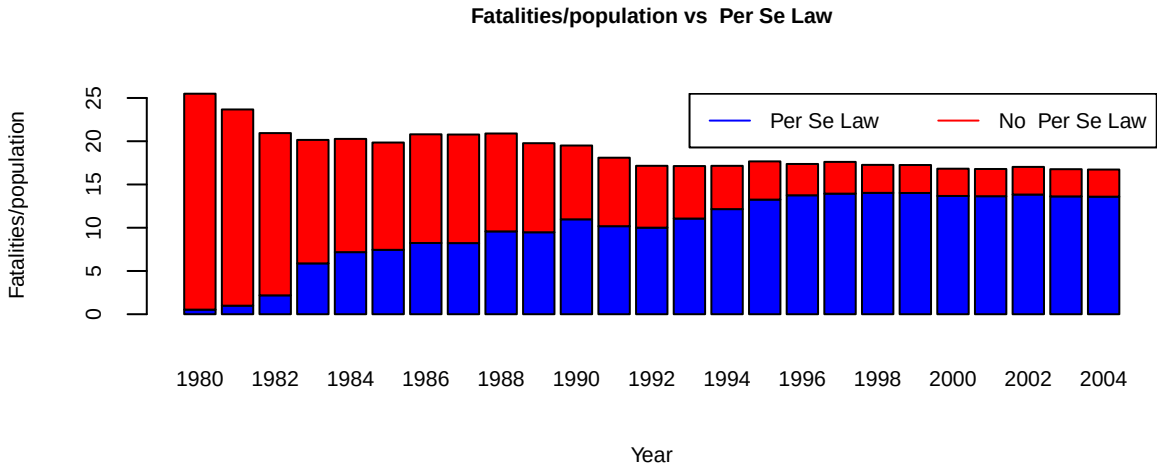
Graduated driver license laws do not appear to impact fatalies very much. Most likely, it will not impact fatalities even accounting for time lags. The changes from graduated license laws from 1999 to 2004 barely impacted fatalities with or without lag effects.

```
plotstackedbox(data, data$gdl, "Gradulated Drivers Law", "GDL")
```



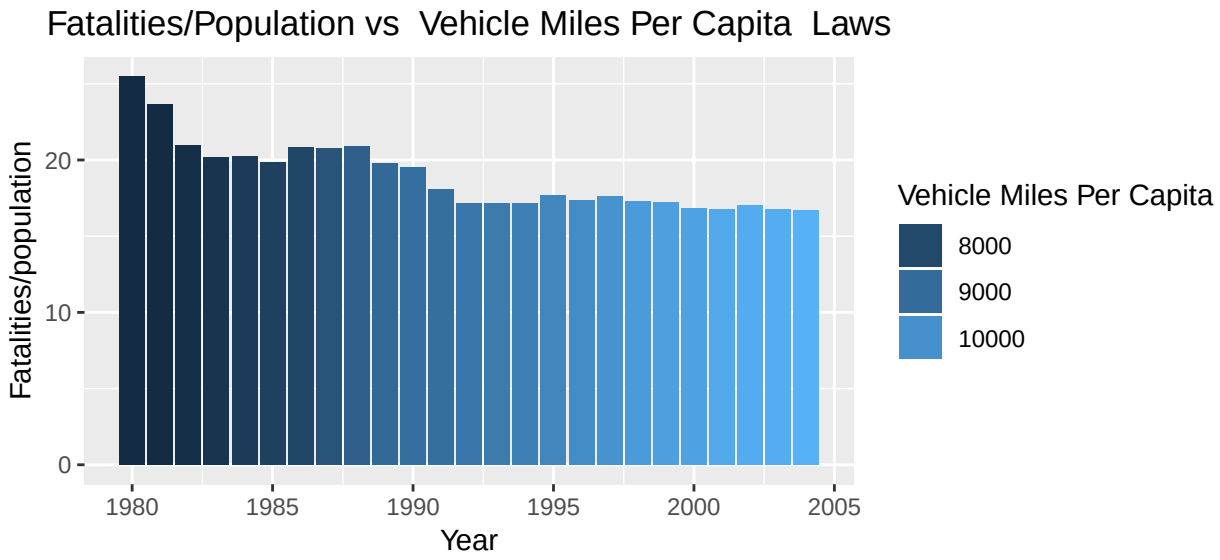
Per se law does not appear to impact fatalities. The increase from 1982 to 1983 did not appear to have a contemporaneous or lagged impact on fatalities. Per se laws may have interactions with BAC laws as it increasese the "harshness" of bac laws.

```
plotstackedbox(data, data$perse, "Per Se Law", "Per Se Law", "topright", 2)
```



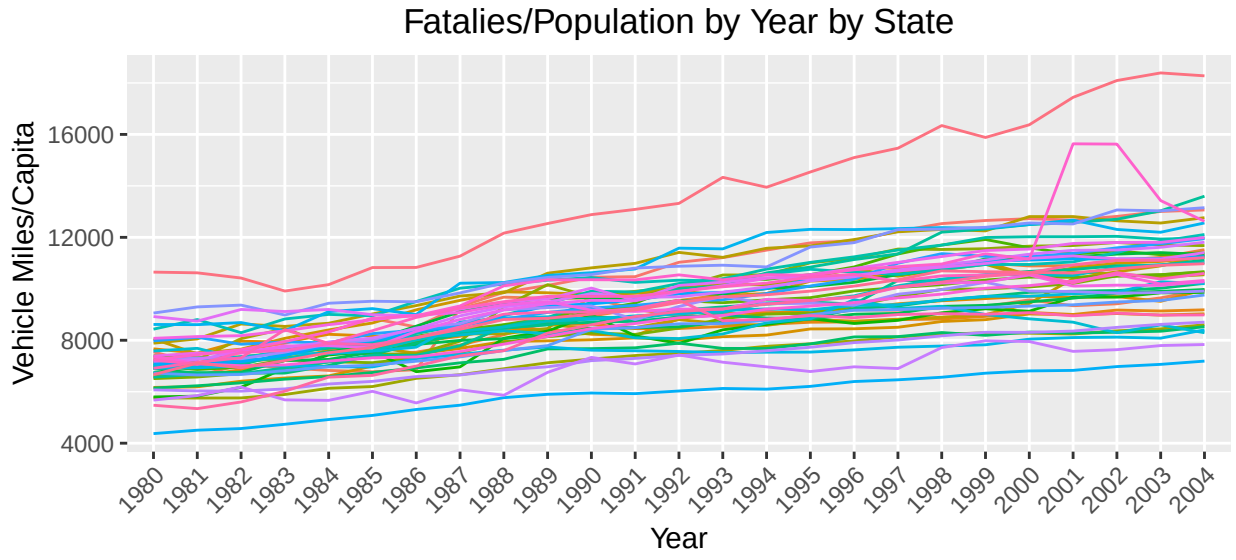
Vehicle miles per capital may be secondarily affected by traffic laws such as graduated license laws. Preliminarily, fatalities appear to decrease as it increases. This makes no sense and is likely to be trending effect through time.

```
plotmix(data, "vehicmilespc", "Vehicle Miles Per Capita")
```



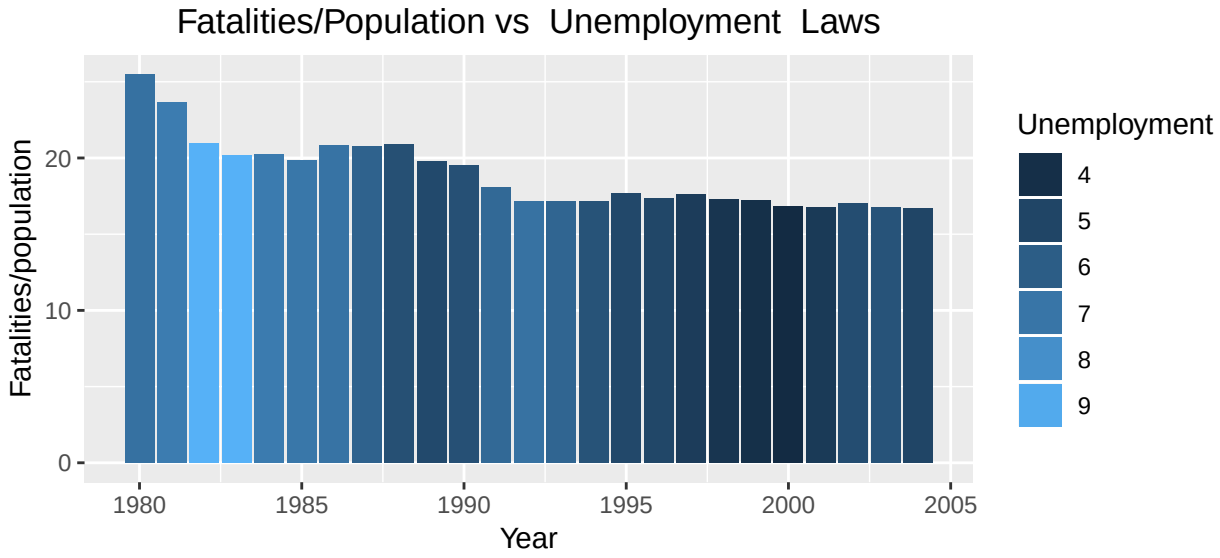
While all states increased **vehicmilespc**, state 46 was interesting in that there was a large increase in 2001-2001 followed by a large drop back down to historical trend by 2003-2004. Closer, state-specific reasoning may be required. Given that we do not have that information, we will not examine it further.

```
ggplot(data, aes(y = vehicmilespc, x = factor(year), group = factor(state), color = factor(state))) +
    geom_line() + theme(axis.text.x = element_text(angle = 45, hjust = 1)) + labs(x = c("Year"),
    y = c("Vehicle Miles/Capita"), title = c("Fatalies/Population by Year by State"),
    color = c("State")) + theme(plot.title = element_text(hjust = 0.5)) + guides(color = FALSE)
```



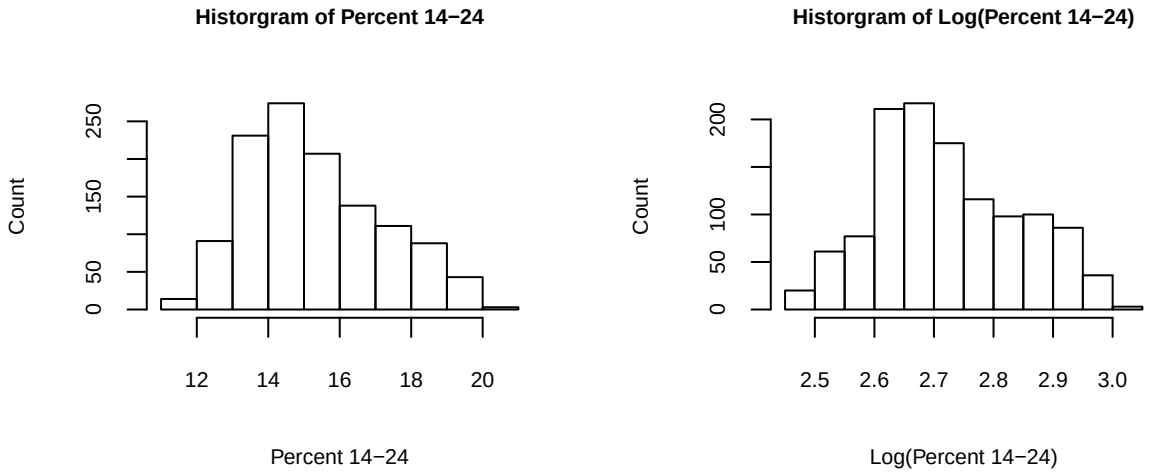
Fatalities do not appear to be affected by unemployment rates at all. In fact, the pattern appears random. We expect this variable to have a β close to 0 in regressions. If we do include the variable, we may want to consider logging it to spread out the values and improve the regression. It does not appear heteroskedastic.

```
plotmix(data, "unem", "Unemployment")
```

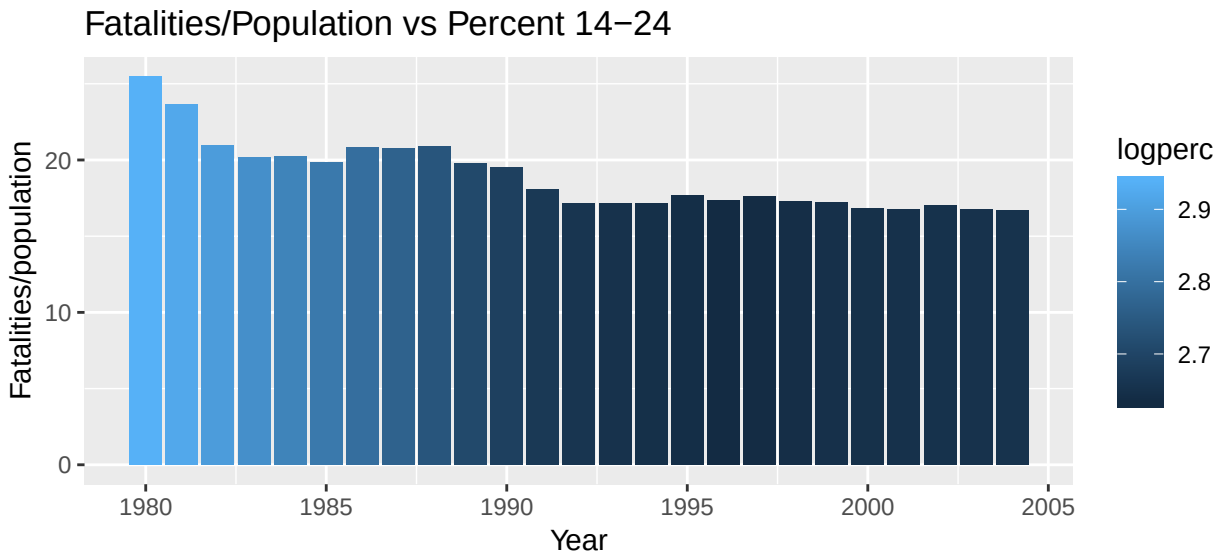


Percent 14-24 is left skewed and a log transformation will be performed. It appears that percent 14-24 is negatively correlated with fatalities. This may potentially be an artifact of the general trend in both population and fatalities.

```
par(mfrow = c(1, 2))
hist(data$perc14_24, main = "Historgram of Percent 14-24", xlab = "Percent 14-24",
    ylab = "Count", cex.main = 0.7, cex.axis = 0.7, cex.lab = 0.7, cex.name = 0.7)
hist(log(data$perc14_24), main = "Historgram of Log(Percent 14-24)", xlab = "Log(Percent 14-24)",
    ylab = "Count", cex.main = 0.7, cex.axis = 0.7, cex.lab = 0.7, cex.name = 0.7)
```

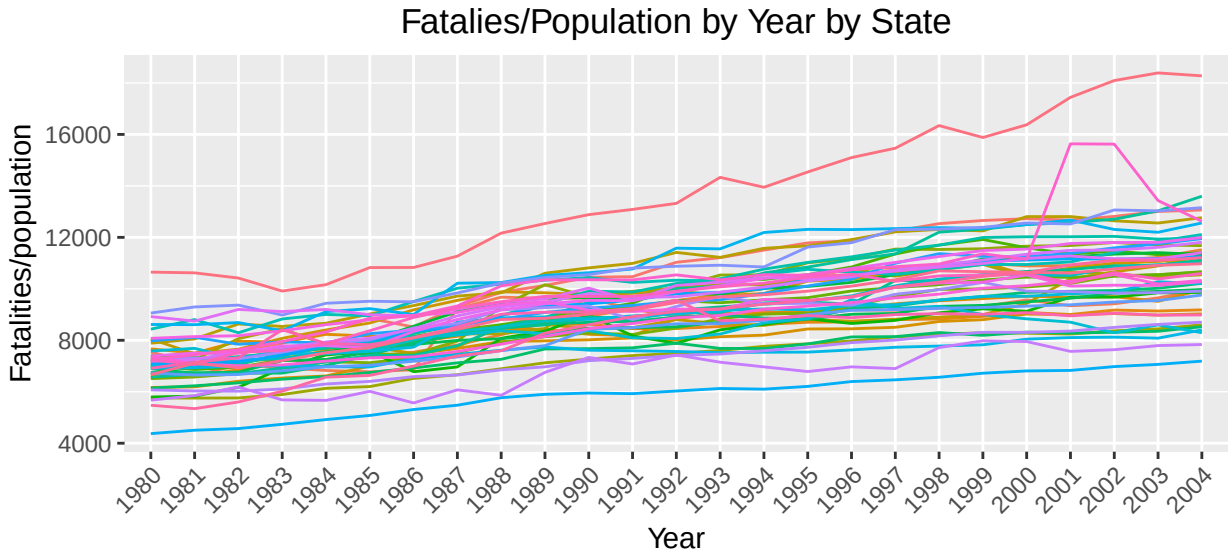



```
tmp = data
tmp$logperc14_24 = log(data$perc14_24)
tmp = data %>% group_by(year) %>% mutate(logperc = log(perc14_24)) %>% summarize_all(funs(mean))
ggplot(tmp, aes(y = totfatrte, x = year, fill = logperc)) + geom_bar(stat = "identity") +
    labs(x = "Year", y = "Fatalities/population", title = "Fatalities/Population vs Percent 14-24")  #+
```



While all states decreased in their population of 14-24 year olds, a few states increased in the ratio. Most significantly, state 45's increase stood out amongst all the states with the sudden jump and decrease between 2001-2003. Again, without more state specific infomration, it's difficult to further examine it.

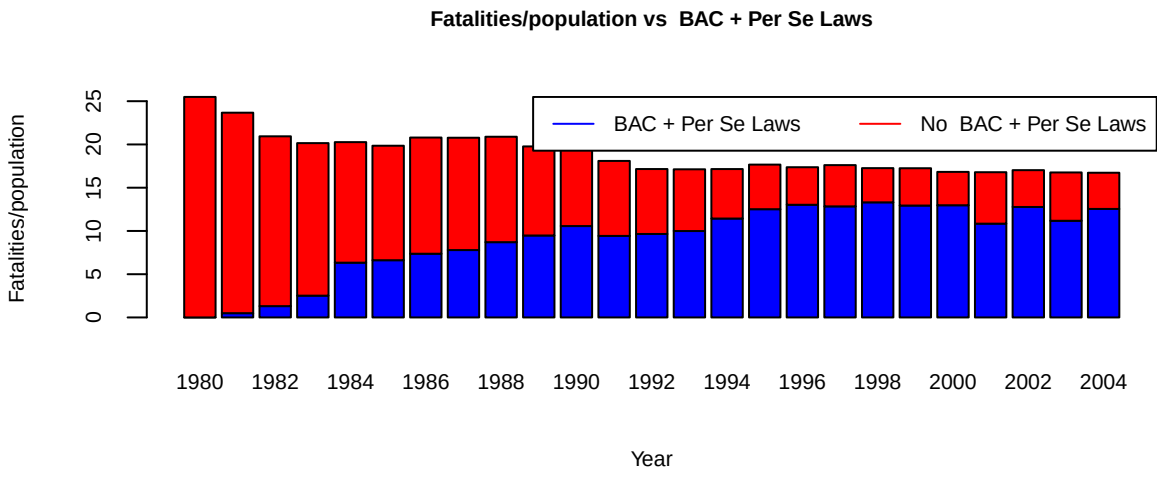
```
ggplot(data, aes(y = vehicmilespc, x = factor(year), group = factor(state), color = factor(state))) +
    geom_line() + theme(axis.text.x = element_text(angle = 45, hjust = 1)) + labs(x = c("Year"),
    y = c("Fatalities/population"), title = c("Fatalies/Population by Year by State"),
    color = c("State")) + theme(plot.title = element_text(hjust = 0.5)) + guides(color = FALSE)
```



Due to the space constraint, bivariate EDA will focus on variables that we have identified as potential interesting interaction variables - `bac`, `perse` and `minage`. We will binarize bac laws to streamline the analysis. BAC + Per Se Law seems to have a time lagged effect on decreasing fatalitie as seen from 1984-1992.

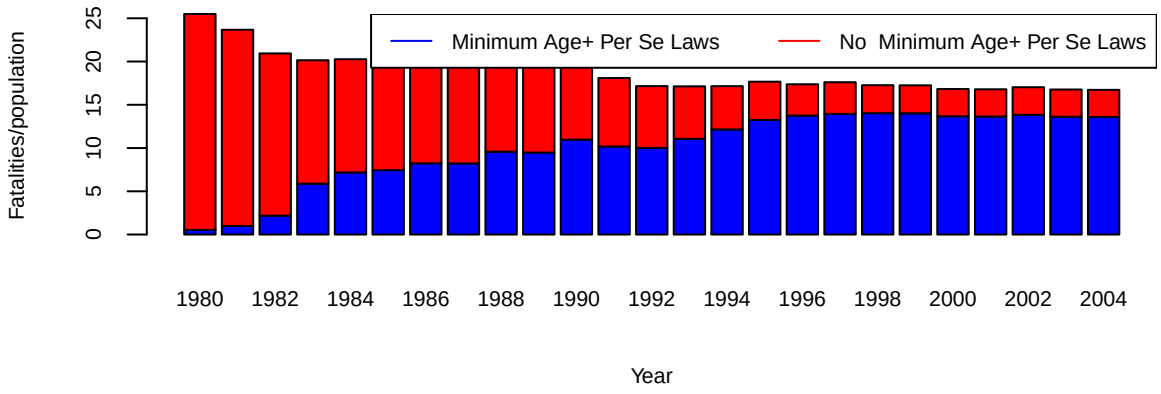
```
tmp = data
tmp$bac = ifelse(tmp$bac08 >= 1 | tmp$bac10 >= 1, 1, 0)
tmp$gdlperse = tmp$gdl * tmp$perse
tmp$bacperse = tmp$bac * tmp$perse
tmp$minageperse = tmp$minage * tmp$perse
tmp$bacminage = tmp$bac * tmp$minage
plotstackedbox(tmp, tmp$bacperse, "BAC + Per Se Laws", "BAC + Per Se Laws", "topright",
```

```
    2)
```



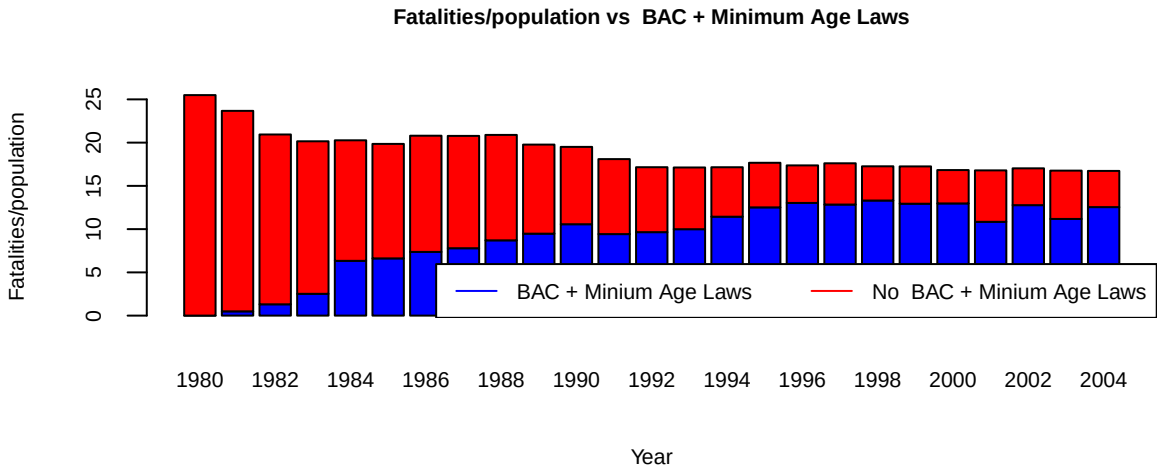
Minimum age and Per Se Laws does not appear to affect fatalities and are unlikely to be a significant variable contemporaneously or lagged.

```
plotstackedbox(tmp, tmp$minageperse, "Minimum Age + Per Se Laws", "Minimum Age+ Per Se Laws",
    "topright", 2)
```



Again, the relationship appears to have a time lagged impact much like BAC + Per Se Laws. However, this may simply be because of the BAC laws. T-tests and time-lagged variables may be done to further the analysis.

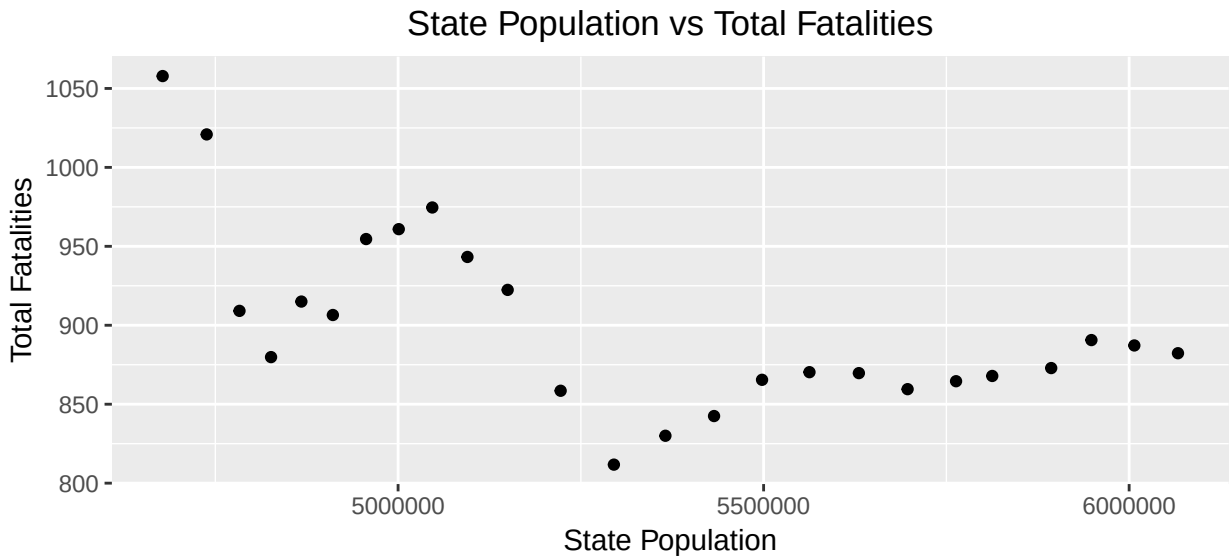
```
plotstackedbox(tmp, tmp$bacperse, "BAC + Minimum Age Laws", "BAC + Minium Age Laws",
    "bottomright", 2)
```



From the EDA, there are variables that appear to be "incorrectly" correlated with fatalities, such as vehicmilespc. Others such as BAC may be better transformed into a binary on-off variable. Per se, zero tolerance, graduated license, minimum drinking age laws appears to have no effect while seatbelts appear to have a contemporaneous impact on fatalites. Speed Limit laws appear to have a lagged effect.

(Question 2) We will first examine the general time trend of fatalities. Recall that our fatalities variable, `totfatrte`, is fatalities per 100,000 population is already normalized by population, so proper analysis of impact of traffic laws on fatalities can be analyzed. Note that traffic laws are most likely uncorrelated with state population as shown in the beginning of the report. Generally, states laws have lower correlation with population suggesting that it has less impact than vehicle miles, which may be affected by state laws. Therefore, total fatalities adjusted by population is probably a better variable to infer state law causalities as noted in the beginning of the report.

```
ggplot(data %>% group_by(year) %>% summarize_all(funs(mean)), aes(y = totfat, x = statepop)) +
    geom_point() + labs(y = "Total Fatalities", x = "State Population", title = "State Population vs To
    theme(plot.title = element_text(hjust = 0.5))
```



SS we can include a chart with state total fatality rate v popultion? that would confirm there's not relationship - LET ME KNOW IF ABOVE WORKS SULLIVAN

```
xtable(cor(data[, c("statepop", "vehicmiles", "minage", "zerotol", "gdl", "seatbelt")]))
```

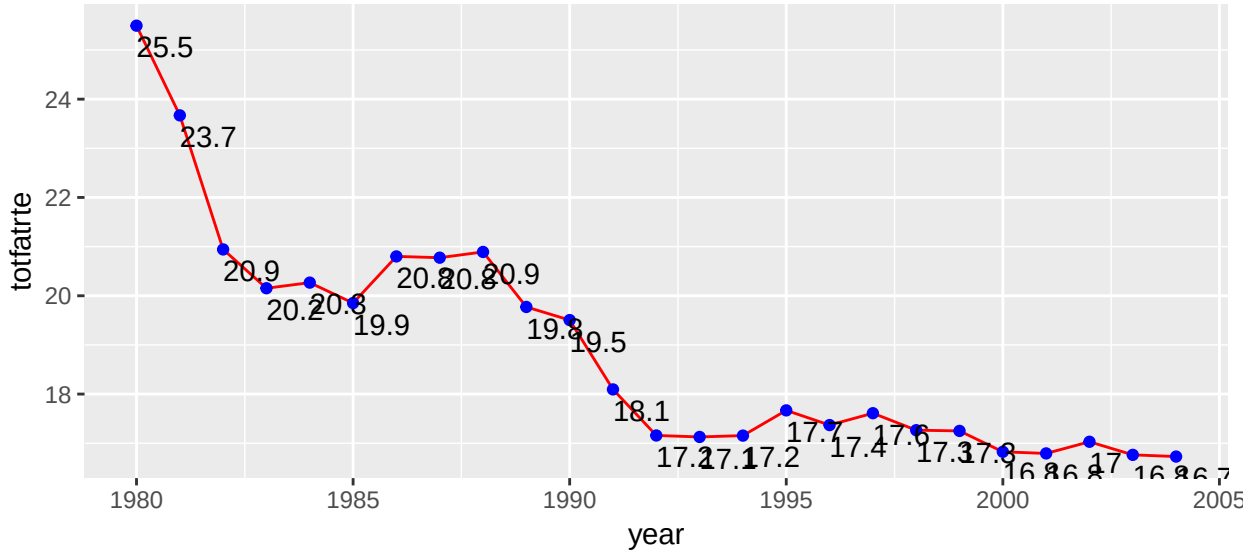
	statepop	vehicmiles	minage	zerotol	gdl	seatbelt
statepop	1.00	0.97	0.10	0.10	0.11	0.04
vehicmiles	0.97	1.00	0.16	0.21	0.20	0.11
minage	0.10	0.16	1.00	0.38	0.20	0.51
zerotol	0.10	0.21	0.38	1.00	0.52	0.46
gdl	0.11	0.20	0.20	0.52	1.00	0.23
seatbelt	0.04	0.11	0.51	0.46	0.23	1.00

A simple plot of the mean `totfatrte` shows that it has decreased over time.

```
tmp = data %>% select(year, totfatrte) %>% group_by(year) %>% summarize_all(funs(mean)) %>%
    as.data.frame
xtable(t(tmp))
```

	1	2	3	4	5	6	7	8	9	10	11
year	1980.00	1981.00	1982.00	1983.00	1984.00	1985.00	1986.00	1987.00	1988.00	1989.00	1990.00
totfatrte	25.49	23.67	20.94	20.15	20.27	19.85	20.80	20.77	20.89	19.77	19.51

```
ggplot(data %>% select(year, totfatrte) %>% group_by(year) %>% summarize_all(funs(mean)),
    aes(year, totfatrte, label = totfatrte)) + geom_line(color = "red") + geom_point(color = "blue") +
    geom_text(aes(label = round(totfatrte, 1)), color = "black", hjust = 0, vjust = 1.5)
```




```
tmp = data %>% select(year, totfatrte) %>% group_by(year) %>% summarise_all(funs(mean))
```

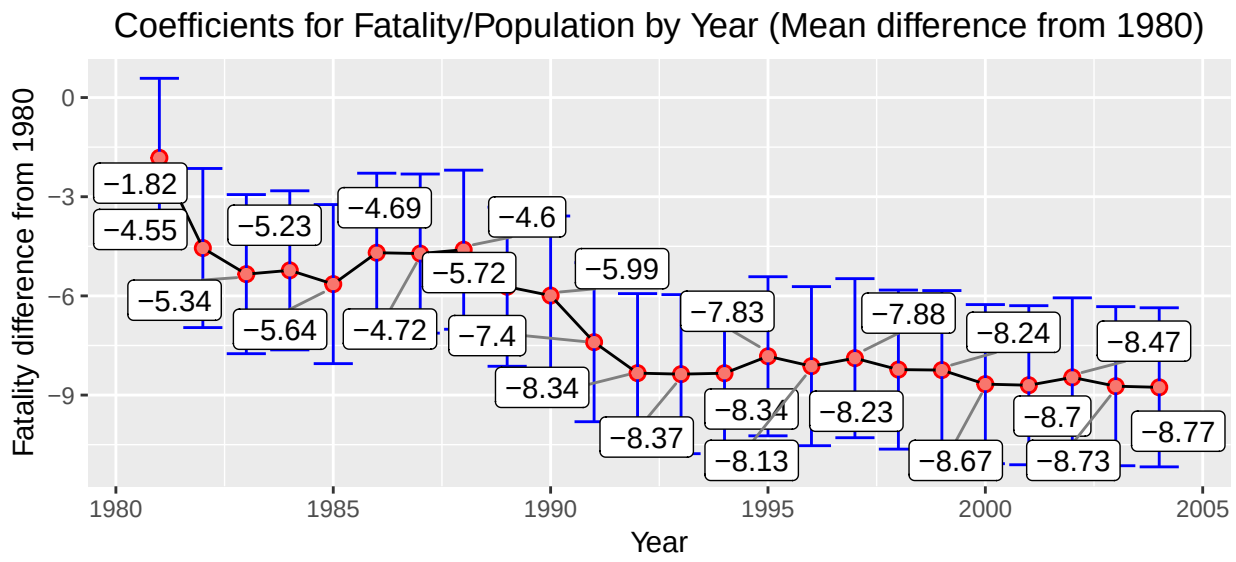
(Question 2) Regression of fatalities vs each year shows that there is a clear significant downward trend. The $F - statistic$ with p-value of ~ 0 shows that the dummy variables for year is jointly significant. We only show the first few coefficients below and the rest are plotted with the standard error. The regression suggests that fatalities have been decreasing through time and the βs show the mean differential between the year t and 1980. The intercept of the regression is the mean fatalities in 1980 and the coefficients is the mean differences from 1980 for each year respectively. Notice the 2 chart are exact same shape after the 1st year (1980). Driving became much safer over time. Note that it's a general trend over time as the +/-95% confidence intervals of βs overlap each other in the surrounding years. We can only conclude that the trend

has decreased but difficult to firmly draw conclusions on any given year.

```
m = lm(totfatrte ~ factor(year), data)
xtable(summary(m)$coefficients[1:5, ])
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	25.49	0.87	29.40	0.00
factor(year)1981	-1.82	1.23	-1.49	0.14
factor(year)1982	-4.55	1.23	-3.71	0.00
factor(year)1983	-5.34	1.23	-4.36	0.00
factor(year)1984	-5.23	1.23	-4.26	0.00

```
tmp = data.frame(confint(m))
tmp = tmp[2:dim(tmp)[1], ]
colnames(tmp) = c("lower", "upper")
tmp$year = 1981:2004
tmp$beta = (tmp$lower + tmp$upper)/2
ggplot(tmp, aes(x = year, y = beta)) + geom_point(size = 2.5, color = "red") + geom_line() +
    geom_errorbar(aes(ymax = tmp$upper, ymin = tmp$lower), color = "blue") + labs(title = "Coefficients
    x = "Year", y = "Fatality difference from 1980") + geom_point(aes(col = "red")) +
    geom_label_repel(aes(label = round(tmp$beta, 2)), box.padding = 0.15, point.padding = 0.5,
        segment.color = "grey50") + guides(color = FALSE) + theme(plot.title = element_text(hjust = 0.5)
```



Note that the residuals are not normally distributed and fails the Shapiro Wilks test. They are also serially correlated as shown by Durbin-Watson test, indicating potentially inefficiency.

```
shapiro.test(m$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  m$residuals
## W = 0.9703, p-value = 5.637e-15
```

```
durbinWatsonTest(m)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.8972667     0.1996691       0
##  Alternative hypothesis: rho != 0
```

(Question 3) We will now expand the previous regression with additional regressors - bac08, bac10, perse, sbprim, sbsecon, sl70plus, gld, perc14_24, unem and vehicmilespc. perc14-24 is logged since it is very left skewed and unem is also logged. Logging spreads out the values for potentially better ergression results. vehicmilepc does not appear to require transformations as it appears more normally/uniformly distributed. The rest of variables are binary and no transformations are done. BAC and speed limit variables are not binarized and no interactions are implemented to limit the report length.

```
library(plm)
data.p = pdata.frame(data, index = c("state", "year"))
m = lm(totfatrte ~ factor(year) + bac08 + bac10 + perse + sbprim + sbsecon + sl70plus +
    gdl + log(perc14_24) + log(unem) + vehicmilespc, data)
xtable(summary(m)$coefficients[26:35, ])
```

	Estimate	Std. Error	t value	Pr(>|t|)
bac08	-2.61	0.53	-4.88	0.00
bac10	-1.45	0.39	-3.69	0.00
perse	-0.54	0.30	-1.81	0.07
sbprim	-0.36	0.49	-0.73	0.47
sbsecon	-0.15	0.43	-0.34	0.73
sl70plus	3.20	0.44	7.19	0.00
gdl	-0.38	0.52	-0.72	0.47
log(perc14_24)	2.73	1.86	1.47	0.14
log(unem)	5.08	0.48	10.57	0.00
vehicmilespc	0.00	0.00	30.85	0.00

We did not not show all the βs for the factor year but they can be found in the R code. bac08 and bac10 have coefficients of -2.605 and -1.453. The β_{bac08} and β_{bac10} represent the impact of having bac08 and bac10 laws in that year (regardless of the year) on the *mean* fatalities across the states. Per se laws also decrease the mean fatalities by -0.535 once it's enacted. Primary seat belt laws does not seem to have an impact on fatalities despite the β_{sbprim} = -0.356. We note that all the β estimators, p-values and other regression statistics above are biased, inconsistent and inefficient as fixed effects and explanatory variable are likely to be correlated. In reality, no conclusion or inference can be made with the estimator since the regression assumptions are violated and estimators are biased. We note that the residuals fail both the Shapiro-Wilks normality test and Durbin-Watson serial correlation test.

```
shapiro.test(m$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  m$residuals
## W = 0.97761, p-value = 1.114e-12
```

```
durbinWatsonTest(m)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.7933425     0.4106811       0
##  Alternative hypothesis: rho != 0
```

Additionally, we can do a pool test to check if the coefficients are consistent vs a fixed effects model. The hypothesis is rejected showing that the estimates are not consistent either (we can take 1 cross-sectional regressional and obtain unbiased estimators *if* unobserved effects are uncorrelated with explanatory variables).

```
pooltest(totfatrte ~ bac08 + bac10 + perse + sbprim + sbsecon + sl70plus + gdl +
    log(perc14_24) + log(unem) + vehicmilespc, data = data.p, model = "within")
```

```
##
##  F statistic
```

```
##
## data:  totfatrte ~ bac08 + bac10 + perse + sbprim + sbsecon + sl70plus +  ...
## F = 3.0655, df1 = 470, df2 = 672, p-value < 2.2e-16
## alternative hypothesis: unstability
```

(Question 4) Given the pooled model assumptions may be violated, we will examine data using a fixed effects model.

```
m.fe = plm(totfatrte ~ bac08 + bac10 + perse + sbprim + sbsecon + sl70plus + gdl +
    log(perc14_24) + log(unem) + vehicmilespc, data = data.p, model = "within")
stargazer(m.fe, no.space = T)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Dec 09, 2018 - 06:39:04 PM

Table 1:

	Dependent variable:
	totfatrte
bac08	-1.841^{***}
	(0.388)
bac10	-1.465^{***}
	(0.269)
perse	-1.666^{***}
	(0.249)
sbprim	-1.725^{***}
	(0.348)
sbsecon	-0.777^{***}
	(0.250)
sl70plus	-1.157^{***}
	(0.247)
gdl	-0.606^{***}
	(0.231)
log(perc14_24)	14.541^{***}
	(1.100)
log(unem)	-3.292^{***}
	(0.323)
vehicmilespc	0.0003^{***}
	(0.0001)
Observations	1,200
R^2	0.536
Adjusted R^2	0.513
F Statistic	132.125^{***} (df = 10; 1142)
Note:	*p<0.1; **p<0.05; ***p<0.01

We note that the residuals are not normal and serially correlated.

```
shapiro.test(m.fe$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  m.fe$residuals
## W = 0.95348, p-value < 2.2e-16
```

```
pbgtest(m.fe, order = 2)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel
##  models
##
## data:  totfatrte ~ bac08 + bac10 + perse + sbprim + sbsecon + sl70plus +     gdl + log(perc14_24) + l
## chisq = 345.82, df = 2, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```
d = data.frame(pooled_coef = summary(m)$coefficients[26:35, 1], pooled_SE = summary(m)$coefficients[26:3
    2], FE_coef = summary(m.fe)$coefficients[, 1], FE_SE = summary(m.fe)$coefficients[,
    2])
xtable(d)
```

	pooled_coef	pooled_SE	FE_coef	FE_SE
bac08	-2.61	0.53	-1.84	0.39
bac10	-1.45	0.39	-1.47	0.27
perse	-0.54	0.30	-1.67	0.25
sbprim	-0.36	0.49	-1.73	0.35
sbsecon	-0.15	0.43	-0.78	0.25
sl70plus	3.20	0.44	-1.16	0.25
gdl	-0.38	0.52	-0.61	0.23
log(perc14_24)	2.73	1.86	14.54	1.10
log(unem)	5.08	0.48	-3.29	0.32
vehicmilespc	0.00	0.00	0.00	0.00

The coefficients are different between the pooled OLS and Fixed Effects regression. The downward bias from pooled OLS due to fixed effects appears in `bac08` and `log(perc14_24)`. `bac10` and `vehicmilespc` is roughly unbiased. `perse`, `gds`, `sbprim`, and `sbsecon` have been biased upwards. More importantly, `sl70plus` and `log(unem)` have both been biased by fixed effects such that the signs of the coefficients are incorrect. No inferences can really be made from the regression as the standard errors are incorrect due to serially correlated residuals.

FE model is better than pooled OLS since it removes the fixed effects. Pooled OLS assumes that there is no correlation between unobserved variable and any of the regressors. This assumption is clearly broken. For example, dry laws, which are unobserved, may be correlated with bac08 laws and affect fatalities. From the pool test and above, we can tell the estimators are biased in the pooled OLS since the estimators are different. Therefore, correlation of fixed effects and explanatory variables definitely exist. The bias causing incorrect signs of coefficients in `sl70plus` and `log(unem)` is notable. For the FE models, the assumption is that the idiosyncratic errors (ϵ_{it}) are uncorrelated conditional on the indepedendent variables and time-invariant unobservable variables. Given the current context, it the FE assumptions are more reasonable as time-invariant variables can be eliminated. Though FE errors are still serially correlated preventing us from doing proper inference, we will at least have unbiased estimators. In the case of pooled OLS, the estimators will not be unbiased and inference cannot be done either.

(Question 5) In comparing FE models with RE models, FE models is likely to be a better estimate in the current context. Like the Pooled OLS, RE model assumes no correlation between fixed effects and explanatory variables. The difference between the pooled OLS and RE models is that RE corrects the serial correlation within the composite error by estimating a correlation. The advantage of RE models over FE is the ability to estimate time-invariant variables. However, it also requires an extremely strong assumption that fixed effects are independent of all explanatory variables across all time periods. Given the endogeneity issues, we believe fixed effects is a much better model than random effects. We can run a Hausman test where the H_0 is that the unique errors are not correlated with the regressors. The H_0 is rejected in the Hausman test.

```
m.re = plm(totfatrte ~ bac08 + bac10 + perse + sbprim + sbsecon + sl70plus + gdl +
    log(perc14_24) + log(unem) + vehicmilespc, data = data.p, model = "random")
phtest(m.fe, m.re)
```

```
Hausman Test
```

data: totfatrte ~ bac08 + bac10 + perse + sbprim + sbsecon + sl70plus + ... chisq = 828.43, df = 10, p-value < 2.2e-16 alternative hypothesis: one model is inconsistent

The final model is

$$\begin{aligned} y_{it} - \bar{y}_i = & -1.95(x_{(bac08)it} - \bar{x}_{bac08}) + -1.56(x_{(bac10)it} - \bar{x}_{bac10}) - 1.56(x_{(perse)it} - \bar{x}_{perse}) \\ & + -1.80(x_{(sbprim)it} - \bar{x}_{sbprim}) + -0.86(x_{(sbsecon)it} - \bar{x}_{sbsecon}) \\ & + -1.12(x_{(sl70plus)it} - \bar{x}_{sl70plus}) + -0.59(x_{(sl70plus)it} - \bar{x}_{sl70plus} \\ & + 14.66(x_{(log(per14_24))it} - \bar{x}_{log(perc14_24)}) + -0.59(x_{(log(unem))it} - \bar{x}_{log(unem)}) \\ & + 0.0003(x_{(vehicmilespc)it} - \bar{x}_{vehicmilespc}) \end{aligned}$$

```
tmp = data %>% group_by(state) %>% summarize_all(funs(mean)) %>% select(state, vehicmilespc) %>%
    as.data.frame
l = NULL
for (i in tmp[, 1]) {
    l = c(l, sd(data[data$state == i, "vehicmilespc"] - tmp[tmp$state == i, "vehicmilespc"]))
}
```

(Question 6) If *vehicmilepc* increase by 1,000 in a time period t assuming the $\bar{vehicmilepc}$ does not change in the FE model, $totfatrte$ is expected to increase 0 with a 95% confidence interval of 0, 0.001. The $\beta_{vehicmilespc}$ may not be an extremely precise estimate. We note that the standard deviation after differencing the mean on vehicmilespc is 474.365 while the standard deviation without differencing is 1825.898. The smaller variation across time causes the estimate to be less precise.

(Question 7) Assuming we have no omitted variable bias (FE model) with autocorrelation and heteroskedasticity exists in errors, the estimators will be inefficient but unbiased. The standard errors estimated on the estimators are incorrect as the variance formula used for estimation is incorrect. It is inefficient as it is no longer the minimum variance estimator.

Exercises:

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

2. How is the our dependent variable of interest *totfatrte* defined? What is the average of this variable in each of the years in the time period covered in this dataset? Estimate a linear regression model of *totfatrte* on a set of dummy variables for the years 1981 through 2004. What does this model explain? Describe what you find in this model. Did driving become safer over this period? Please provide a detailed explanation.

```
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)

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)

Coefficients: (Intercept) d81 d82 d83 d84
25.495 -1.824 -4.552 -5.342 -5.227
d85 d86 d87 d88 d89
-5.643 -4.694 -4.720 -4.603 -5.722
d90 d91 d92 d93 d94
-5.989 -7.400 -8.337 -8.367 -8.339
d95 d96 d97 d98 d99
-7.826 -8.125 -7.884 -8.229 -8.244
d00 d01 d02 d03 d04
-8.669 -8.702 -8.465 -8.731 -8.766

3. Expand your model in *Exercise 2* by adding variables *bac08, bac10, perse, sbprim, sbsecon, sl70plus, gdl, perc14_24, unem, vehicmilespc*, and perhaps *transformations of some or all of these variables*. Please explain carefully your rationale, which should be based on your EDA, behind any transformation you made. If no transformation is made, explain why transformation is not needed. How are the variables *bac8* and *bac10* defined? Interpret the coefficients on *bac8* and *bac10*. Do *per se laws* have a negative effect on the fatality rate? What about having a primary seat belt law? (Note that if a law was enacted sometime within a year the fraction of the year is recorded in place of the zero-one indicator.)

```
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 + sl70plus + gdl + perc14_24 +
    unem + vehicmilespc, data)

lm2
```

Call: lm(formula = totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89 + d90 + d91 + d92 + d93 + d94 + d95 + d96 + d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04 + bac08 + bac10 + perse + sbprim + sbsecon + sl70plus + gdl + perc14_24 + unem + vehicmilespc, data = data)

Coefficients: (Intercept) d81 d82 d83 d84
-2.716054 -2.175479 -6.595970 -7.396690 -5.850394
d85 d86 d87 d88 d89
-6.483252 -5.852796 -6.367393 -6.591578 -8.070967
d90 d91 d92 d93 d94
-8.958670 -11.068552 -12.878398 -12.730718 -12.364833
d95 d96 d97 d98 d99
-11.952549 -13.876377 -14.258378 -15.041676 -15.090547
d00 d01 d02 d03 d04
-15.443946 -16.183715 -16.724350 -17.021308 -16.711273
bac08 bac10 perse sbprim sbsecon
-2.498483 -1.417565 -0.620108 -0.075335 0.067280
sl70plus gdl perc14_24 unem vehicmilespc
3.347914 -0.426911 0.141590 0.757053 0.002925

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

```
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 + sl70plus + gdl + perc14_24 + unem +
    vehicmilespc, data)
```

Model Formula: totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89 + d90 + d91 + d92 + d93 + d94 + d95 + d96 + d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04 + bac08 + bac10 + perse +

sbprim + sbsecon + sl70plus + gdl + perc14_24 + unem + vehicmilespc

Coefficients: bac08 bac10 perse sbprim sbsecon -2.4984831 -1.4175652 -0.6201081 -0.0753347 0.0672804 sl70plus gdl perc14_24 unem vehicmilespc 3.3479143 -0.4269107 0.1415903 0.7570529 0.0029254

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

Fixed effects are constant across the states and this may most likely not true, considering our EDA. Random effects vary across the states, and this is more likely to be close to reality.

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

7. If there is serial correlation or heteroskedasticity in the idiosyncratic errors of the model, what would be the consequences on the estimators and their standard errors?