Statistical Methods for Discrete Response, Time Series, and Panel Data (W271): Lab 3

Group 5

Justin Trobec **Ajit Barhate** **David Linnard Wheeler**

U.S. traffic fatalities: 1980-2004

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

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

Part 1

(30%) 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.*

Load libraries (code hidden for brevity)

Load data

```
# Import
load("driving.RData")
# Rename
df <- data
# Structure
str(df)
```

```
# Summary
summary(df)
```

- After loading and inspecting the structure and dimensions of the data, we see that it is comprised of 1200 x 56 data frame. Moreover, from the `str(df)` output, we see the `class` of each variable- this will be important information to revisit as we progress with EDA.
- Now that we understand the basic data structure, we inspect the contents of the data:

```
# Summary
describe(df)
```

- From the summaries above (results are hidden to comply with the page limit) we see that the data does not appear to have any missing values (confirmed below) and that scales and summary statistics vary among the variables, except the indicator variables like `d80`.
- Also from the summaries above we can see that variables like `slnone`, `sl55`, `sl65`, `sl70`, `sl70plus`, `sl75`, `zeroto`, `per se`, `minage`, `bac08`, and `bac10` are not just comprised of 0 and 1- they have fractional values expressed as $\frac{m}{12}$ where m is the month in which the law was enacted. We will not transform these variables because (i) they contain information and (ii) to binarize these variables would be assume that ther law enactments that occur within the year have the *same* effect as those laws that were enacted on say January 1st. That is to say that the 12-m month has no effect on `totfatrte`.
- We will verify the records per year and state are symmetric; the following commands verify that we have 48 records for each year and 25 for each state. Tabular results are hidden for brevity.

```
table(df$state)
table(df$year)
```

- After inspection of the data structure and summaries, we will proceed with more EDA.

Explortory Data Analysis

- Before exploration of the univariate and multivariate relationships among the variables with tabular and graphical methods, we first check for missing data.

```
# Are there any missing values/NAs?
df[!complete.cases(df),]
```

- We found no missing values.
- Before we proceed with the univariate analysis section below we will look for evidence of top and bottom-coded variables.

```
head(df);tail(df)
```

- Evidence of no top or bottom-coded variables was not detected (i) when we look at the top and bottom of the dataframe or (ii) when we inspect histograms for sharp thresholds (see appendix). Again, the results are hidden to comply with the page limit.
- Next, we examine `totfatrte` over time for each state. Note that we log-transformed `totfatrte` to achieve normality. This transformation will be defended below.

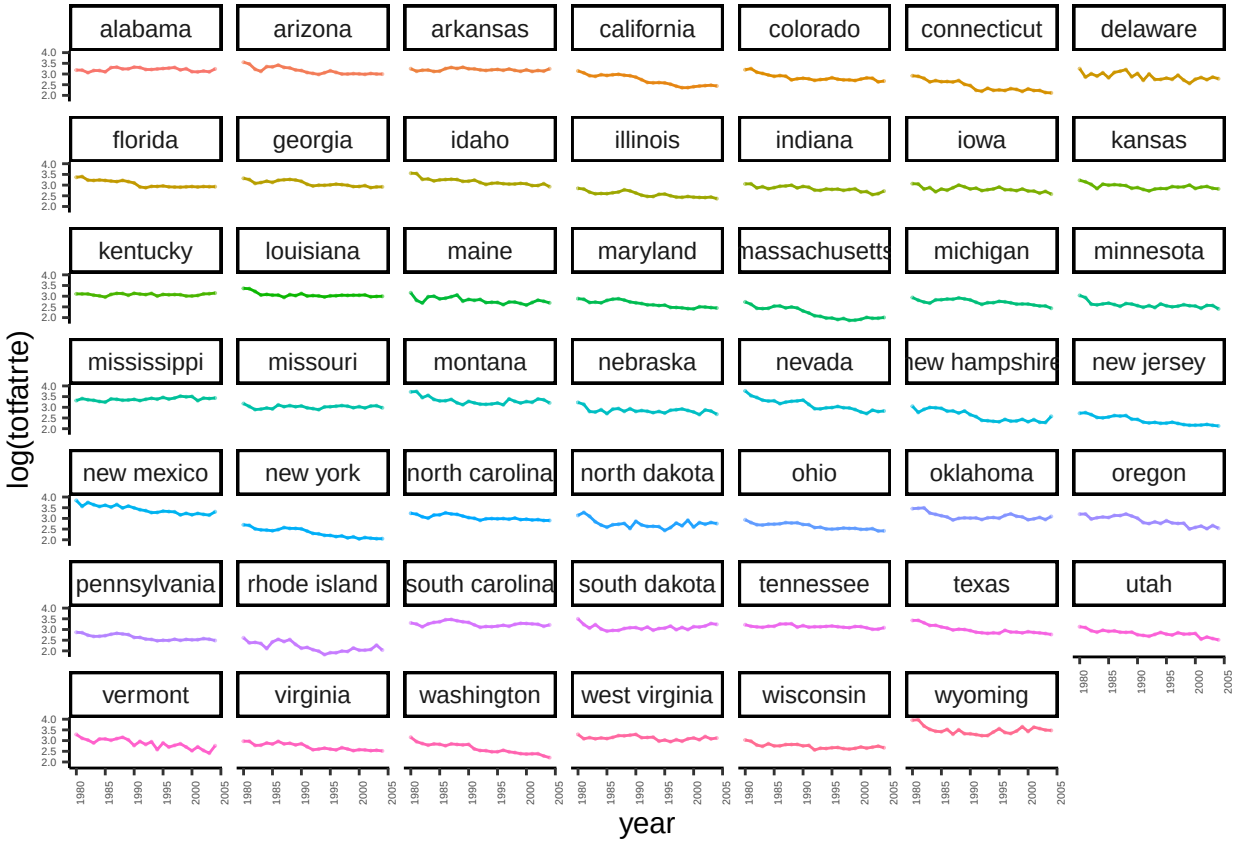


Figure 1: Growth curves of log(total fatalities per 100,000 population) over time for each state

- From the growth curves above, we see some variation in log(`totfatrte`) for each state over time. Some states, like Arkansas, display flat rates, while others, like Wyoming have fallen only to rise again. This plot helps us understand `totfatrte` within states but does not help us understand regional differences in `totfatrte`.
- To further explore (i) the distribution of `totfatrte` and (ii) regional differences in `totfatrte`, we now plot the average `totfatrte` for each state.
- From the histogram presented on the right side of the yearly box-plots below, we see that `totfatrte` is normally distributed, after log-transformation. When we plot histograms of `totfatret` for each state we see scale shifts in the distribution of `totfatrte` between states.

For example, the distribution of `totfatrte` in Wyoming is shifted towards higher values compared to the distribution in other states like Massachusetts.

- From the map below we can see some general patterns. First, we see that some western and southern states like Wyoming, New Mexico, and Mississippi, display relatively high values of log(`totfatrte`). Conversely, some northeastern states, like Massachusetts and New York display relatively low values of log(`totfatrte`). This figure tell us a lot about regional fatalities but not about fatality rates over time.
- To get a better sense of nationwide fatality rates over time, we next show `totfatrte` over time for all states combined.

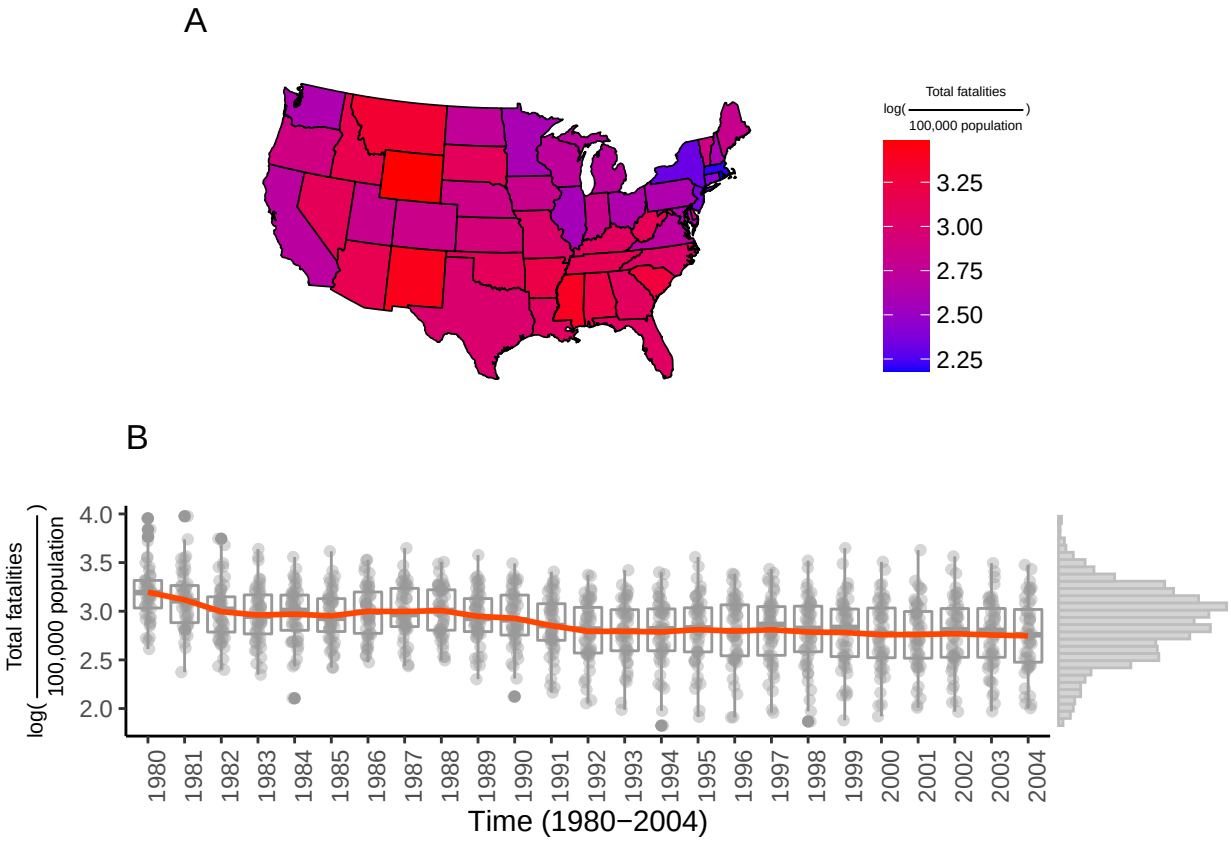


Figure 2: Nationwide log(total fatality rates per 100,000 population)

- The plot above shows the slow drop in `totfatrte` nationwide over time. First, `totfatrte` decreases from 1980 to the early 1980's. This decrease in `totfatrte` also corresponds with the inception and actions of Mothers Against Drunk Driving. During the mid to late-1980's `totfatrte` increases. Starting in the mid-1990s `totfatrte` remains stable over time.
- What is responsible for these changes in `totfatrte`? One reasonably possibility is that `vehicmilespc` contributes to `totfatrte`. The more miles people drive, the higher the probability of an accident. A time plot that shows `vehicmilespc` over time and a scatterplot that show `vehicmilespc` vs `totfatrte` are presented in the appendix. Since `totfatrte` is

positively correlated with and potentially confounded by `vehicmilespc`, it will be an important covariate in the models we use in this lab. Likewise, other demographic variables like `unem` and `perc14_24` will be potentially important covariates to include in the models. These variables and their relationship with `totfatrte` are explored in the appendix.

- Another possibility, not mutually exclusive from the first, is that the introduction, adoption, and enforcement of the below laws contribute to `totfatrte`.

Variable	Law
	Age
`zerotol` & `gdl`	- zero tolerance & graduated drivers license
	Speed
`slnone`, `sl55`, `sl65`, `sl70`, `sl70plus`,& `sl75`	- speed limit: not defined, 55, 65, 70, 70+, and 75
	Seatbelt
`sbprim` & `sbsecon`	- primary and secondary
	Alcohol
`zerotol`, `per se`, `minage`, `bac08`, & `bac10`	- zero tolerance, administrative license revocation , minimum drinking age, blood alcohol limit .08 & 0.1

- Evidence of these relationships is presented below, on page 6.

- In figure 3 we see some interesting patterns in law adoption over time. For example, as `bac10` and `sl65` rise and fall, `bac08` is adopted by more and more states over time. Like `bac10`, `gdl` and `zerotol` rise sharply in the early 1990s. Other laws, like the `perse`, `sbprim`, and `slnone` (not show to save space) rise more slowly and consistently over time. Three of the speed limit laws, `sl70`, `sl70plus`, and `sl75` rise abruptly in 1995 and plateau thereafter.
- In all, we observed various interesting relationships between the dependent variable, `totfatrte`, and putative predictors in the EDA. Next, we will elucidate these relationships with pooled OLS, fixed and random effects models.

Part 2

(15%) How is the our dependent variable of interest *totfatrte* defined? What is the average of this variable in each of the years in the time period covered in this dataset? Estimate a linear regression model of *totfatrte* on a set of dummy variables for the years 1981 through 2004. What does this model explain? Describe what you find in this model. Did driving become safer over this period? Please provide a detailed explanation.

- `totfatrte` is the "total fatalities per 100,000 population". These data are collected annually from every state. We log-transformed this variable to achieve normality.
- The average of `totfatrte` in each of the years, across states, is presented below:

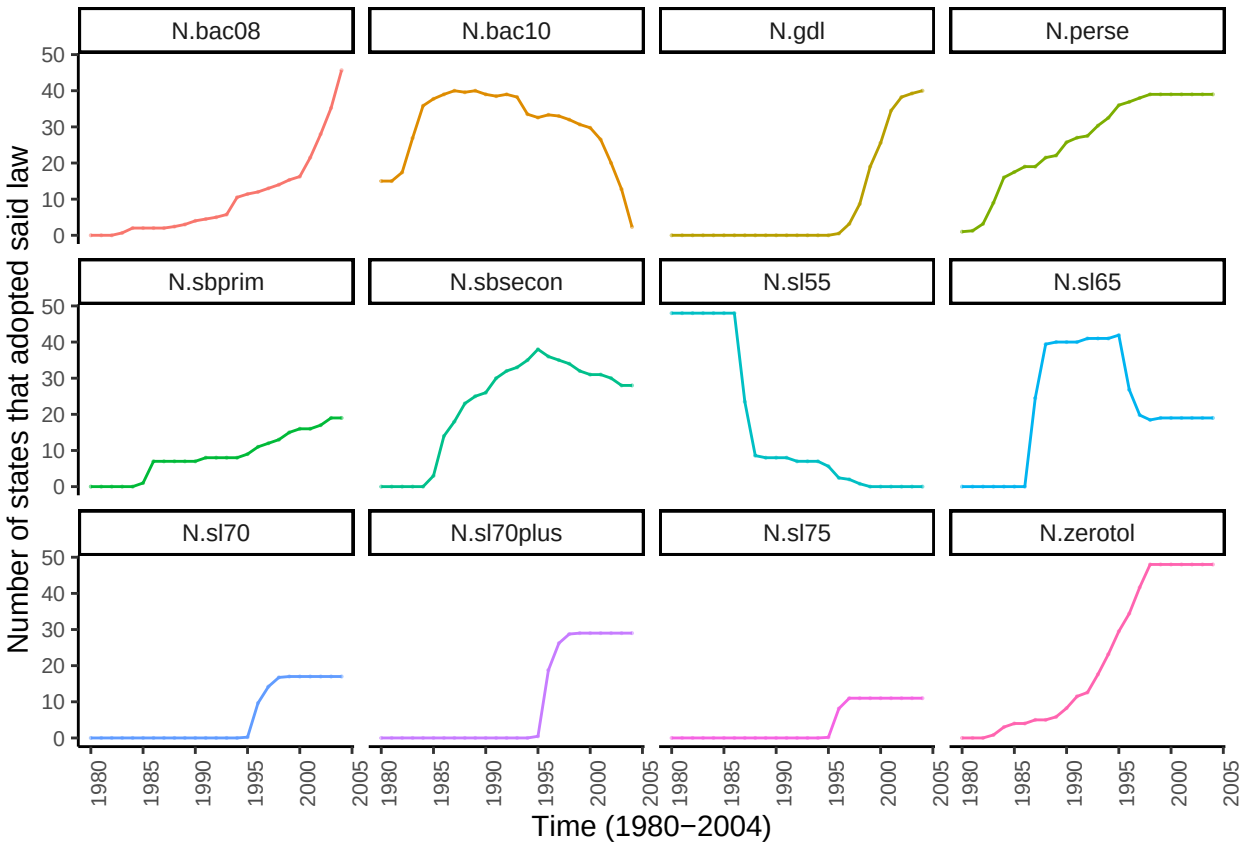


Figure 3: The number of states that adpted each law over time

```r
# Plot average growth over time
ftl.per.yr <- df %>%
  select(year, totfatrte) %>%
 group_by(year) %>%
  summarise("year.mu" = mean(totfatrte))
```

	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991
Avg. Fat. Rte	25.49	23.67	20.94	20.15	20.27	19.85	20.8	20.77	20.89	19.77	19.51	18.09

	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Avg. Fat. Rte	17.16	17.13	17.16	17.67	17.37	17.61	17.27	17.25	16.83	16.79	17.03	16.76	16.73

- Note that the fluctuations over time described by the above table are also captured by the figure below.

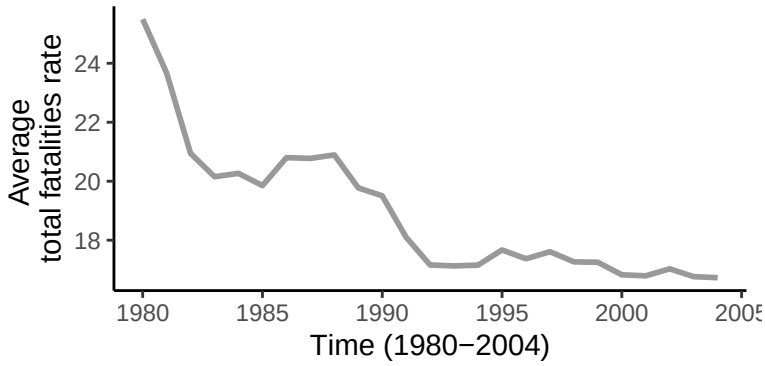


Figure 4: Average total fatalities rate per year

- As above and in figure 1 from Freeman 2007 we can see the drop in average nationwide fatalities over time.
- Now we will estimate a pooled linear regression model with log(`totfatrte`) on a set of dummy variables for the years 1981 through 2004. Note that we will not include the year 1980 and it will be the base level of the model.

```r
# Coerce data into pdata.frame
panel.df <- pdata.frame(df, index=c("state","year"))
# Estimate model
pOLS.mod.2 <- plm(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,
              model="pooling",
              data=panel.df)
```

```
summary(pOLS.mod.2)
# Robust standard errors
sqrt(diag(vcovHC(pOLS.mod.2, type='HC0')))
# Summary
coeftest(pOLS.mod.2, vcov = vcovHC, type = "HC0")
```

- The general mathematical form of the pooled OLS model with time indicators only is below. To save space, the coefficients for this model are included when we display all the coefficients in answer to Part 4. To see the specific coefficient values, look ahead to that table. Finally, note that the robust standard errors are used to accommodate the heteroscedastic residuals seen in the appendix. Non-robnust standard errors produce different standard errors and p-values.

$$log(totfatrre) = \beta_0 + \beta_1 d_{81} + \beta_2 d_{82} ... \beta_{24} d_{04}$$

- A summary table of all three models is presented, via stargazer, in response to question 4.
- The above model is a pooled OLS model that explains the average *totfatrte* across all 48 continental states and how it changes over years.
- The intercept is the average fatality rate of all states for the year 1980 (base average `totfatrte`). Every coefficient for the dummy variable for each year then explains how the average `totfatrte` changes relative to the base level of 1980.
- We can see that all coefficients of the dummy variables are negative which means the average `totfatrte`of all states is decreasing relative to the average `totfatrte` in 1980. We can also validate this visually from Figure 4 above.
- Note that all coefficients of dummy variables except for `d81` are highly statistically significant since $p < \alpha = 0.05$. However, we note that the coefficient for dummy variable `d81` becomes highly statistically significant when we use heteroskedasticity robust standard errors.
- The adjusted R^2 tells us that the model only explains about 11% of the total variance in the `totfatrte`.
- Note that the coefficients of dummy variables for each year are negative with higher descent in the early years and staying almost flat with very low fluctuations from the year 1992. Based on this we can conclude that the model tells us that the `totfatrte` has been lower than the 1980 over years and has been staying almost flat since 1992. This is also visible from the time series plot of average `totfatrte`. Given this is a pooled OLS model which violates the basic independence assumption, the estimates from the model will not be reliable. Hence, although there is visual evidence and this model supports it, we can not reliably say that the driving has gotten safe over years.

Part 3

(15%) 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 *bac08* 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.)

- Before we expand the model to include the variables recommended above, we note that, for the reasons addressed above and summarized below, we will only transform the response variable `totfatrte` and explanatory variable `vehicmilespc`. Both are log-transformed below. The former is transformed to achieve normality and the latter to pull the tails of the distribution towards the center. For the "binary" indicator variables that are expressed as some decimal form of $\frac{m}{12}$ we do not transform the data so as to retain all information that would be lost if we assumed the monthly contributions of law enactments were negligible and rounded to 0 or 1. In other words, we are not transforming the data to (i) preserve information and (ii) not to assume that the fractional monthly components during which the given law was enacted had the same effect as say 0 if $\frac{m}{12} < 0.5$ or 1 if $\frac{m}{12} > 0.5$.
- `bac08` and `bac10` are defined as the blood alcohol concentration limit of 0.08 or 0.10, respectively. States without these laws for a given year are assigned 0 . Otherwise, states are assigned $\frac{m}{12}$ where m is the month of law enactment.
- Now we estimate the expanded pooled OLS model:

```
# Estimate model
pOLS.mod.3 <- plm(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 + bac08 + bac10 + perse + sbprim +
             sbsecon + sl70plus + gdl + perc14_24 + unem +
             log(vehicmilespc),
             model="pooling",
             data=panel.df)

#Summarize model
summary(pOLS.mod.3)

# Robust standard errors
sqrt(diag(vcovHC(pOLS.mod.3, type='HC0')))
coeftest(pOLS.mod.3, vcov = vcovHC, type = "HC0")
```

- The mathematical form of the model is described below. As mentioned in Part 2 above, the full set of specific coefficient values can be found in the table in answer to Question 4.

$$\begin{aligned} log(totfatrre) =& \beta_0 + \beta_1 d_{81} + \beta_2 d_{82} ... \beta_{24} d_{04} + \beta_{25} bac_{08} + \beta_{26} bac_{10} \\ &+ \beta_{27} perse + \beta_{29} sbprim + \beta_{30} sbsecon + \beta_{31} sl70plus \\ &+ \beta_{32} gdl + \beta_{33} perc14\_24 + \beta_{34} unem + \beta_{35} log(vehicmilespc) \end{aligned}$$

- The coefficients for `bac08` and `bac10` are not statistically significant.

  1) `bac08` = -0.057667; robust standard error = 0.085126; p-value > 0.05.
  2) `bac10` = -0.016215; robust standard error = 0.069030 ; p-value > 0.05.

- Thus, if we ignore the assumption of independence, which we have violated by pooling the data, we can interpret the coefficients for `bac08` and `bac10` as follows. For each additional state that adopts `bac08` laws, there is a 5.8% reduction in `totfatrte` while holding other explanatory variables constant.

- Similarly, for each additional state that adopts `bac10` laws, there is a 1.6% reduction in `totfatrte` while holding other explanatory variables constant.

- Thus, the `bac08` law appears to be *more* effective than `bac10` if we ignore the violated independence assumption.

- The coefficient for `perse` is -0.025041. The negative sign of the coefficient indicates that it has a negative effect on the `totfatrte`. However, given its p-value $> \alpha = 0.05$, it is not statistically significant in this model. The interpretation of the coefficient is that for each additional state adopting the `perse` law there is a 2.5% reduction in `totfatrte` while holding other explanatory variables constant.

- The coefficient for primary seat belt law (`sbprim`) is 0.014182. The positive sign of the coefficient indicates that this law surprisingly has positive effect on `totfatrte`. However, given its p-value $> \alpha = 0.05$, this explanatory variable is not statistically significant in this model. The interpretation of the coefficient is that for each additional state adopting the primary seat belt law there is a 1.4% increase in `totfatrte` while holding other explanatory variables constant. Given that we violate the independence assumption, we have to be cautious with this interpretation (especially, the directionality on its effect).

- Just to clarify, the interpretations above assume that we have not violated the independence assumption. Since we have violated this assumption by ignoring the correlation of errors among cross-sectional units and pooling the data (as opposed to sampling 1200/48=25 observations), the coefficients, standard errors, and p-values should be interpreted with caution. Moreover, we assume, by using pooled OLS that the composite error $\alpha_i + \epsilon_{it}$ is not correlated with the observed explanatory variables. This is a strong assumption!

Part 4

(15%) 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?

- Below we estimate a fixed effects model.

```
# Estimate model
FE.mod.4 <- plm(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 + bac08 + bac10 + perse + sbprim +
            sbsecon + sl70plus + gdl + perc14_24 + unem + log(vehicmilespc),
            model="within", data=panel.df)

# Summarize model
summary(FE.mod.4)

# Robust standard errors
sqrt(diag(vcovHC(FE.mod.4, type='HC3',method="arellano")))
coeftest(FE.mod.4, vcov = vcovHC, type = "HC3",method="arellano")
```

- One thing interesting to note is that this model specification ends up deriving the same coefficients as a model that omits the dummy variables, and uses the `effect='twoways'` parameter. For example, the following model produces the same coefficients for bac_{08}, bac_{10}, *perse*, *sbprim*, *sbsecon*, *sl70plus*, *gdl*, *perc*14_24, *unem*, *and vehicmilespc*, without including all the dummy variables in the specification. Because other questions include the time dummy variables, we will stick with that approach, and not use the model below, even though it is equivalent and less verbose. Similarly, we could use `factor(year)` to include the dummy variables for each year.

```
FE.no.time <- plm(log(totfatrte) ~ bac08 + bac10 + perse + sbprim +
            sbsecon + sl70plus + gdl + perc14_24 + unem + log(vehicmilespc),
            model="within", effect="twoways", data=panel.df)
```

- The following table lists the coefficients for the pooled and fixed-effects models we've built up till now:

```
stargazer(pOLS.mod.2, pOLS.mod.3, FE.mod.4,
          se=list(sqrt(diag(vcovHC(pOLS.mod.2, type='HC0'))),
                  sqrt(diag(vcovHC(pOLS.mod.3, type='HC0'))),
                  sqrt(diag(vcovHC(FE.mod.4, type='HC3',method="arellano")))),
          single.row = TRUE,model.names = FALSE,model.numbers= FALSE,
          digits = 2,dep.var.caption  = "log of total fatality rate per 100,000 population", de
          column.labels = c("pooled OLS (Q2)", "pooled OLS (expanded/Q3)", "fixed effect model
```

- Compare coefficients from each model

Table 4:

	log of total fatality rate per 100,000 population		
	pooled OLS (Q2)	pooled OLS (expanded/Q3)	fixed effect model
d81	-0.08^{***} (0.02)	-0.10^{***} (0.02)	-0.06^{***} (0.02)
d82	-0.20^{***} (0.02)	-0.32^{***} (0.03)	-0.12^{***} (0.02)
d83	-0.24^{***} (0.02)	-0.37^{***} (0.05)	-0.15^{***} (0.03)
d84	-0.23^{***} (0.02)	-0.31^{***} (0.05)	-0.20^{***} (0.02)
d85	-0.24^{***} (0.02)	-0.34^{***} (0.06)	-0.22^{***} (0.03)
d86	-0.20^{***} (0.02)	-0.33^{***} (0.08)	-0.18^{***} (0.04)
d87	-0.20^{***} (0.03)	-0.37^{***} (0.09)	-0.22^{***} (0.04)
d88	-0.19^{***} (0.02)	-0.38^{***} (0.10)	-0.25^{***} (0.05)
d89	-0.25^{***} (0.02)	-0.47^{***} (0.12)	-0.32^{***} (0.06)
d90	-0.27^{***} (0.02)	-0.52^{***} (0.12)	-0.33^{***} (0.06)
d91	-0.34^{***} (0.03)	-0.64^{***} (0.13)	-0.37^{***} (0.07)
d92	-0.40^{***} (0.03)	-0.75^{***} (0.14)	-0.43^{***} (0.07)
d93	-0.40^{***} (0.03)	-0.74^{***} (0.14)	-0.45^{***} (0.07)
d94	-0.41^{***} (0.03)	-0.73^{***} (0.15)	-0.48^{***} (0.07)
d95	-0.38^{***} (0.03)	-0.71^{***} (0.15)	-0.48^{***} (0.08)
d96	-0.40^{***} (0.03)	-0.83^{***} (0.17)	-0.52^{***} (0.08)
d97	-0.39^{***} (0.03)	-0.86^{***} (0.17)	-0.54^{***} (0.08)
d98	-0.41^{***} (0.03)	-0.91^{***} (0.17)	-0.59^{***} (0.08)
d99	-0.41^{***} (0.03)	-0.92^{***} (0.17)	-0.60^{***} (0.09)
d00	-0.44^{***} (0.03)	-0.93^{***} (0.18)	-0.63^{***} (0.09)
d01	-0.44^{***} (0.03)	-0.97^{***} (0.18)	-0.61^{***} (0.09)
d02	-0.43^{***} (0.03)	-1.00^{***} (0.18)	-0.59^{***} (0.09)
d03	-0.44^{***} (0.03)	-1.02^{***} (0.18)	-0.59^{***} (0.09)
d04	-0.45^{***} (0.03)	-1.01^{***} (0.19)	-0.62^{***} (0.09)
bac08		-0.06 (0.09)	-0.02 (0.03)
bac10		-0.02 (0.07)	-0.02 (0.02)
perse		-0.03 (0.05)	-0.06^{***} (0.02)
sbprim		0.01 (0.06)	-0.04^{*} (0.02)
sbsecon		0.03 (0.04)	0.004 (0.02)
sl70plus		0.24^{***} (0.05)	0.07^{***} (0.02)
gdl		-0.03 (0.05)	-0.02 (0.02)
perc14_24		0.02 (0.02)	0.02^{*} (0.01)
unem		0.04^{***} (0.01)	-0.03^{***} (0.005)
log(vehicmilespc)		1.54^{***} (0.14)	0.66^{***} (0.14)
Constant	3.20^{***} (0.04)	-11.07^{***} (1.39)	
Observations	1,200	1,200	1,200
R^2	0.13	0.67	0.72
Adjusted R^2	0.11	0.66	0.70
F Statistic	7.06^{***} (df = 24; 1175)	68.84^{***} (df = 34; 1165)	86.65^{***} (df = 34; 1118)

Note: *p<0.1; **p<0.05; ***p<0.01

```r
data.frame("pooled.OLS" = c(pOLS.mod.3$coefficients['bac08'],
                            pOLS.mod.3$coefficients['bac10'],
                            pOLS.mod.3$coefficients['perse'],
                            pOLS.mod.3$coefficients['sbprim']),
           "FixedEffectModel" = c(FE.mod.4$coefficients['bac08'],
                            FE.mod.4$coefficients['bac10'],
                            FE.mod.4$coefficients['perse'],
                            FE.mod.4$coefficients['sbprim']))
```

```
##         pooled.OLS FixedEffectModel
## bac08  -0.05766704      -0.02239447
## bac10  -0.01621546      -0.02020523
## perse  -0.02504052      -0.05732876
## sbprim  0.01418214      -0.04289376
```

- Below are observations for from the above table:

    1) `bac08`: The directionality of effect on `totfatrte` in the fixed effect model remains the same as it is in the pooled OLS model. However, the effect on `totfatrte` reduces to less than half (from 5.8% to 2.2% reduction in `totfatrte` for every additional state adopting this law).

    2) `bac10`: The directionality of effect on `totfatrte` in the fixed effect model remains the same as it is in the pooled OLS model. The effect of adoption of this law now comes a lot closer to `bac08`.

    3) `perse`: The directionality of effect on `totfatrte` in the fixed effect model remains the same as it is in the pooled OLS model. However, the effect on `totfatrte` increases to more than double (from 2.5% to 5.7% reduction in `totfatrte` for every additional state adopting this law).

    4) `sbprim`: Interestingly, the directionality of effect on `totfatrte` in the fixed effect model changes to that of the pooled OLS model. The directionality now makes more sense practically given this law will practically reduce the number of fatalities in case of accidents. The effect of `sbprim` on `totfatrte` is now negative and with a higher magnitude compared to the pooled OLS model. For each additional state adopting this law now is expected to see about 4.3% reduction in `totfatrte` keeping all other variables constant.

- The estimates from the fixed effect model are likely more reliable because since we are violating fewer assumptions.

- For both models we assume the typical OLS assumptions. Of these the assumption of no perfect multicollinearity (see appendix) and normality appear to satisfied. Assumptions that are likely violated are discussed below.

- For the pooled OLS model(s) we assume the typical OLS assumptions. Thus, we are violating the assumption of independence by pooling the data for cross-sectional units/states across

time. To resolve this problem, we could have subsetted a year of data and proceeded with OLS. Moreover, we assume, that the composite error $\alpha_i + \epsilon_{it}$ is not correlated with the observed explanatory variables, x_{it} - this is a very strong assumption that, if violated, can introduce heterogenity bias, as described by Wooldridge. We also assume that there are no idiosyncratic qualities of each state within each year and no fixed effects over time. If these assumptions are violated than the estimators are biased and inconsistent (Wooldridge).

- For the fixed effect model, we assume OLS assumptions and that ϵ_{it} does not vary stochastically over $_i$ or $_t$. We assume that there are idiosyncratic qualities of each state but. if these attributes do not change over time, they are soaked up by the fixed effects.

- Thus, based on the discussion above, we argue that the fixed effects estimates are more reliable. More explicitly, the fixed effects model estimates are more reliable than the pooled OLS estimates because:

- For the pooled OLS estimates, we assume:

  - The data are independent and identically distributed. This assumption is clearly violated since we sampled data from 48 states over 25 years. If we just sampled a single year cross-section, we could argue otherwise.
  - The composite error $\alpha_i + \epsilon_{it}$ is not correlated with the observed explanatory variables. The assumption seems unreasonable for reasons formulated below.
  - There are no idiosyncratic feature of each state and no fixed effects over time. This assumption also seems unreasonable for reasons discussed above and below.

- For the fixed effect model, we assume:

  - ϵ_{it} does not vary stochastically over $_i$ or $_t$. The assumption seems unreasonable for reasons discussed above and below.
  - strict exogeneity for the explanatory variables. If the idiosyncratic error u_{it} is uncorrelated with the explanatory variable across all time periods, x_{ijt} than the fixed effects estimator is not biased (Wooldridge).

- Finally, before we proceed with question 5 we test the Lagrange multiplier H_0 that the pooled OLS model is better than a fixed effects model

```
# null: pooled OLS better than fixed
pFtest(FE.mod.4, pOLS.mod.3)
```

```
##
##  F test for individual effects
##
## data:  log(totfatrte) ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 +  ...
## F = 104.38, df1 = 47, df2 = 1118, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

- We rejected the H_0 and conclude that the fixed effect model, for the many reasons described above, is a better approach than the pooled OLS model.

Part 5

(10%) Would you perfer to use a random effects model instead of the fixed effects model you built in *Exercise 4*? Please explain.

- First we fit a random-effects model:

```
# Fit Random effects model
RE.mod.5 <- plm(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 + bac08 + bac10 + perse + sbprim +
            sbsecon + sl70plus + gdl + perc14_24 + unem + log(vehicmilespc),
            model="random", data=panel.df)
```

```
# Robust standard errors
summary(RE.mod.5)
sqrt(diag(vcovHC(FE.mod.4, type='HC3',method="white2")))
coeftest(FE.mod.4, vcov = vcovHC, type = "HC3",method="white2")
```

- The results of the expanded pooled OLS from part 3, the fixed effects model from part 4, and the random effects model from part 5 are present in table 5 below.

- To test the H_0 that the idiosyncratic errors, u_{it} are not correlated with the regressors, and the random effects model is preferred, we use the Hausman test:

```
# Hausman test
phtest(FE.mod.4, RE.mod.5)
```

```
##
##  Hausman Test
##
## data:  log(totfatrte) ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 +  ...
## chisq = 74.906, df = 34, p-value = 6.586e-05
## alternative hypothesis: one model is inconsistent
```

- In this case, the Hausman test tells us to reject the null hypothesis, meaning that the unique errors are correlated, and we should prefer a fixed effects model to the random effects model.
- Our intuition corroborates the results of the Hausman test since the cross sectional units, the $\frac{48}{50}$ contiguous US states, are not a random selection of states from a universe of many possible states. Instead they are a very specific subset of the US states- those that are contiguous. Thus, we should used a fixed effects model.

Table 5:

	log of total fatality rate per 100,000 population		
	pooled effect model (Q3)	fixed effect model (Q4)	random effect model (Q5)
d81	−0.10*** (0.02)	−0.06*** (0.02)	−0.06*** (0.02)
d82	−0.32*** (0.03)	−0.12*** (0.02)	−0.13*** (0.02)
d83	−0.37*** (0.05)	−0.15*** (0.03)	−0.16*** (0.02)
d84	−0.31*** (0.05)	−0.20*** (0.02)	−0.20*** (0.02)
d85	−0.34*** (0.06)	−0.22*** (0.03)	−0.23*** (0.02)
d86	−0.33*** (0.08)	−0.18*** (0.04)	−0.19*** (0.03)
d87	−0.37*** (0.09)	−0.22*** (0.04)	−0.23*** (0.03)
d88	−0.38*** (0.10)	−0.25*** (0.05)	−0.26*** (0.03)
d89	−0.47*** (0.12)	−0.32*** (0.06)	−0.34*** (0.03)
d90	−0.52*** (0.12)	−0.33*** (0.06)	−0.35*** (0.03)
d91	−0.64*** (0.13)	−0.37*** (0.07)	−0.39*** (0.03)
d92	−0.75*** (0.14)	−0.43*** (0.07)	−0.46*** (0.04)
d93	−0.74*** (0.14)	−0.45*** (0.07)	−0.47*** (0.04)
d94	−0.73*** (0.15)	−0.48*** (0.07)	−0.50*** (0.04)
d95	−0.71*** (0.15)	−0.48*** (0.08)	−0.50*** (0.04)
d96	−0.83*** (0.17)	−0.52*** (0.08)	−0.55*** (0.04)
d97	−0.86*** (0.17)	−0.54*** (0.08)	−0.57*** (0.04)
d98	−0.91*** (0.17)	−0.59*** (0.08)	−0.62*** (0.04)
d99	−0.92*** (0.17)	−0.60*** (0.09)	−0.63*** (0.04)
d00	−0.93*** (0.18)	−0.63*** (0.09)	−0.66*** (0.04)
d01	−0.97*** (0.18)	−0.61*** (0.09)	−0.64*** (0.04)
d02	−1.00*** (0.18)	−0.59*** (0.09)	−0.62*** (0.04)
d03	−1.02*** (0.18)	−0.59*** (0.09)	−0.62*** (0.04)
d04	−1.01*** (0.19)	−0.62*** (0.09)	−0.66*** (0.05)
bac08	−0.06 (0.09)	−0.02 (0.03)	−0.03 (0.02)
bac10	−0.02 (0.07)	−0.02 (0.02)	−0.02* (0.01)
perse	−0.03 (0.05)	−0.06*** (0.02)	−0.06*** (0.01)
sbprim	0.01 (0.06)	−0.04* (0.02)	−0.04*** (0.02)
sbsecon	0.03 (0.04)	0.004 (0.02)	0.005 (0.01)
sl70plus	0.24*** (0.05)	0.07*** (0.02)	0.08*** (0.01)
gdl	−0.03 (0.05)	−0.02 (0.02)	−0.02* (0.01)
perc14_24	0.02 (0.02)	0.02* (0.01)	0.02*** (0.005)
unem	0.04*** (0.01)	−0.03*** (0.005)	−0.02*** (0.003)
log(vehicmilespc)	1.54*** (0.14)	0.66*** (0.14)	0.74*** (0.06)
Constant	−11.07*** (1.39)		−3.63*** (0.51)
Observations	1,200	1,200	1,200
R^2	0.67	0.72	0.71
Adjusted R^2	0.66	0.70	0.70
F Statistic	68.84*** (df = 34; 1165)	86.65*** (df = 34; 1118)	2,851.99***
Note:			*p<0.1; **p<0.05; ***p<0.01

Part 6

(10%) 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.

- Given that we have log-log transformation of `totfatrte` and `vehicmilespc` in the fixed effects model, the interpretation of the coefficient for `vehicmilespc` is the elasticity of `totfatrte` with respect to `vehicmilespc`. Let's assume the coefficient of the log of `vehicmilespc` is β in the fixed effects model. The interpretation of this coefficient is that every 1% increase in `vehicmilespc` will result in the increase of β% in the dependent variable, `totalfatrte`.
- Mathematically, the interpretation is represented as below.

$$\%\triangle \text{ totfatrte} = \%\triangle \text{ vehicmilespc} \times \beta$$

- Note that β is constant for all levels of `vehicmilespc`.
- The effect of increase in the number of miles driven per capita (`vehicmilespc`) on `totfatrte` is dependent on its percentage change.
- Let's take a couple of examples. In the first case, the `vehicmilespc` for a given state is 10,000 and in the second case it is 15,000. We are interested to know the effect of increasing this by 1,000 miles on the `totfatrte`. Below, we calculate this effect.

```
# Estimated effect of 1000 miles increase in vehicmilespc
beta <- FE.mod.4$coefficients['log(vehicmilespc)']
data.frame(base_vehicmilespc = c('10,000', '15,000'), beta=beta, increase_in_miles=1000, pct_e
```

```
##   base_vehicmilespc      beta increase_in_miles pct_effect_on_totfatrte
## 1            10,000 0.6585186              1000                6.585186
## 2            15,000 0.6585186              1000                4.390124
```

- Thus, *ceteris paribus*, if the number of miles driven per capita (`vehicmilespc`) increase by 1,000 miles from 10,000 then the total fatalities rate (`totfatrte`) increases by 6.59%.
- Similarly, *ceteris paribus*, if the number of miles driven per capita (`vehicmilespc`) increase by 1,000 miles from 15,000 then the total fatalities rate (`totfatrte`) increases by 4.39%.
- The effect seems practical and reasonable because as people drive more, there are more chances of getting into accidents and thus for increase in fatalities. As expected, the effect is more pronounced when the percentage change in `vehicmilespc` is higher.

Part 7

(5%) 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?

- The estimators will be biased and not efficient (Wooldridge) if the idiosyncratic errors, u_{it} are correlated.
- To test the H_0^1 that there is not serial correlation in the errors we used the Breusch-Godfrey/Wooldridge test.
- To test the H_0^2 that that the residuals across states are not correlated we used the Bruesh-Pagan Lagrange Multiplier test for independence.
- To test the H_0^3 that the residuals are homoskedasticity we used the Bruesch-Pagan test for homoskedasticity.

```
# Breusch-Godfrey/Wooldridge test for serial correlation
pbgtest(FE.mod.4)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel models
##
## data:  log(totfatrte) ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 +     d89 + d90 + d91
## chisq = 256.21, df = 25, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```
# cross-sectional dependence
pcdtest(FE.mod.4,
        test = c("lm"))
```

```
##
##  Breusch-Pagan LM test for cross-sectional dependence in panels
##
## data:  log(totfatrte) ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 +     d89 + d90 + d91
## chisq = 2853.9, df = 1128, p-value < 2.2e-16
## alternative hypothesis: cross-sectional dependence
```

```
# Bruesh-Pagan test for homoskedasticity
bptest(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 + bac08 + bac10 + perse + sbprim +
            sbsecon + sl70plus + gdl + perc14_24 + unem + log(vehicmilespc),
            data=panel.df, studentize=F)
```

```
##
##  Breusch-Pagan test
##
## data:  log(totfatrte) ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 +     d89 + d90 + d91
## BP = 90.939, df = 34, p-value = 4.333e-07
```

- We rejected H_0^1: that there is no serial correlation in the errors; H_0^2: that the residuals are not correlated across cross-sectional units, states; and H_0^3 that the residuals are homoskedasticity.
- Thus, since the idiosynratic errors are serially correlated and heteroskedasticity, the estimators and their standard errors will be biased and inefficient, respectively.
- Finally, we synthesize our response to the question: **"Do changes in traffic laws affect traffic fatalities?"**. First, we note that we cannot make any casual claims here even though we have controlled for various factors, some of which are potential confounders. Second, we echo the conclusions of Freeman (2007) who, armed with similar data that analyzed differently, converged on the same conclusion: the revision of blood-alcohol content (BAC) laws from 0.1 g/dL(`bac10`) to 0.08 g/dL (`bac08`) does not appear to affect total traffic fatilities **once** we include covariates that control for potential confounders, like `vehiclemilespc`.Conversely, other laws, like the the administrative license revocation (i.e. per se law,`perse`), the primary seatbelt law (`sbprim`), and the 70 mile plus per hour speed limit law significantly affected total traffic fatalities.
- These conclusions, that the administrative license revocation, primary seatbelt,and the 70 mile plus per hour speed limit laws are consistent with the expectations of traffic safety experts expressed here: https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/810878.pdf. Moreover, the effects of these laws on traffic fatalities corroborates our intuitions that (i) seat belts (ii) immediate revocation of drivers licenses, and (iii) slower speed limits all reduce fatalities by keeping people safe in their vehicles and removing drunk drivers from the roadways.

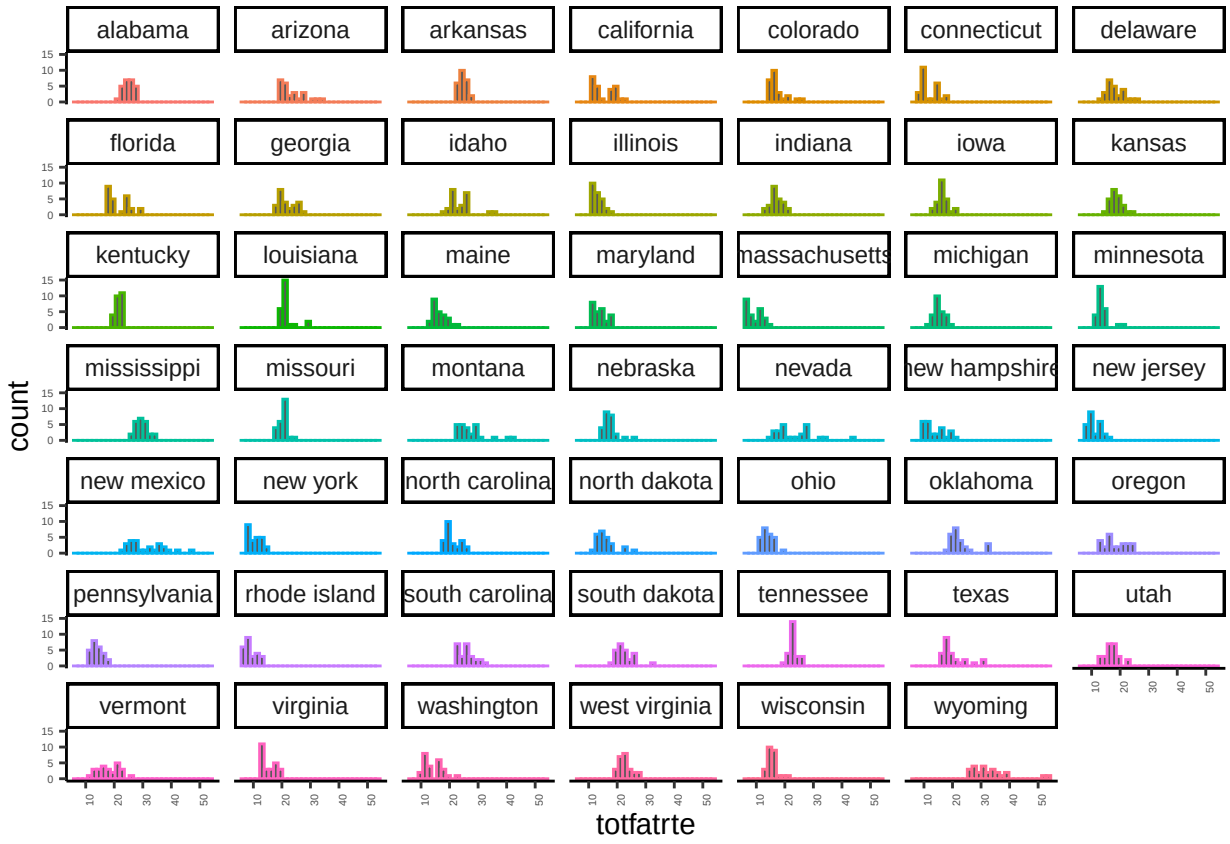
Appendix

Question 1 (Additional EDA)

- Evidence of top- or bottom-coded values of `totfatrte` are not present for any state since no sharp thresholds beyond which values are censored are apparent.

```
# By state
ggplot(Statesdf, aes(totfatrte,colour = region)) +
  geom_histogram() +
  facet_wrap(~region) +
  theme_classic() +
  theme(legend.position="none",
      axis.text.y=element_text(size=4),
      axis.text.x=element_text(size=4, angle = 90))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



- `vehiclemilespc` vs time and `vehiclemilespc` vs `totfatrte`.

- Relationships between `totfatrte` and `perc14_24` and `unem`.

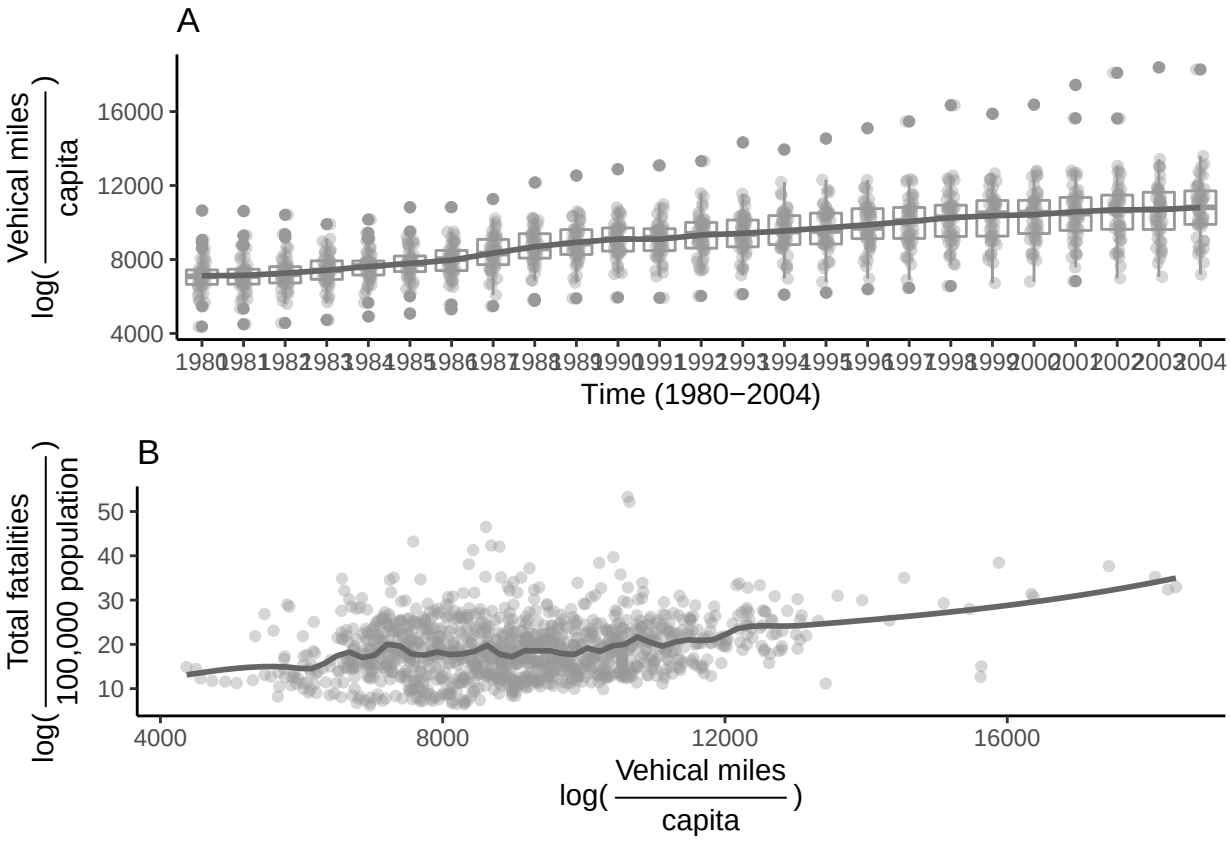


Figure 5: Relationship of vehicle miles per capita with total fatalities per 100,000 population

```
# Scatterplot matrix
GGally::ggpairs(df,
        aes(alpha=1/4),
        mapping=ggplot2::aes(colour = factor(state),alpha=1/4),
        columns=c("totfatrte","perc14_24", "unem" ,"vehicmilespc", "year"),
        lower= list(continuous = "points"),
        upper=list(continuous="density"),
        diag=list(continuous="densityDiag"))+
        theme_classic()
```

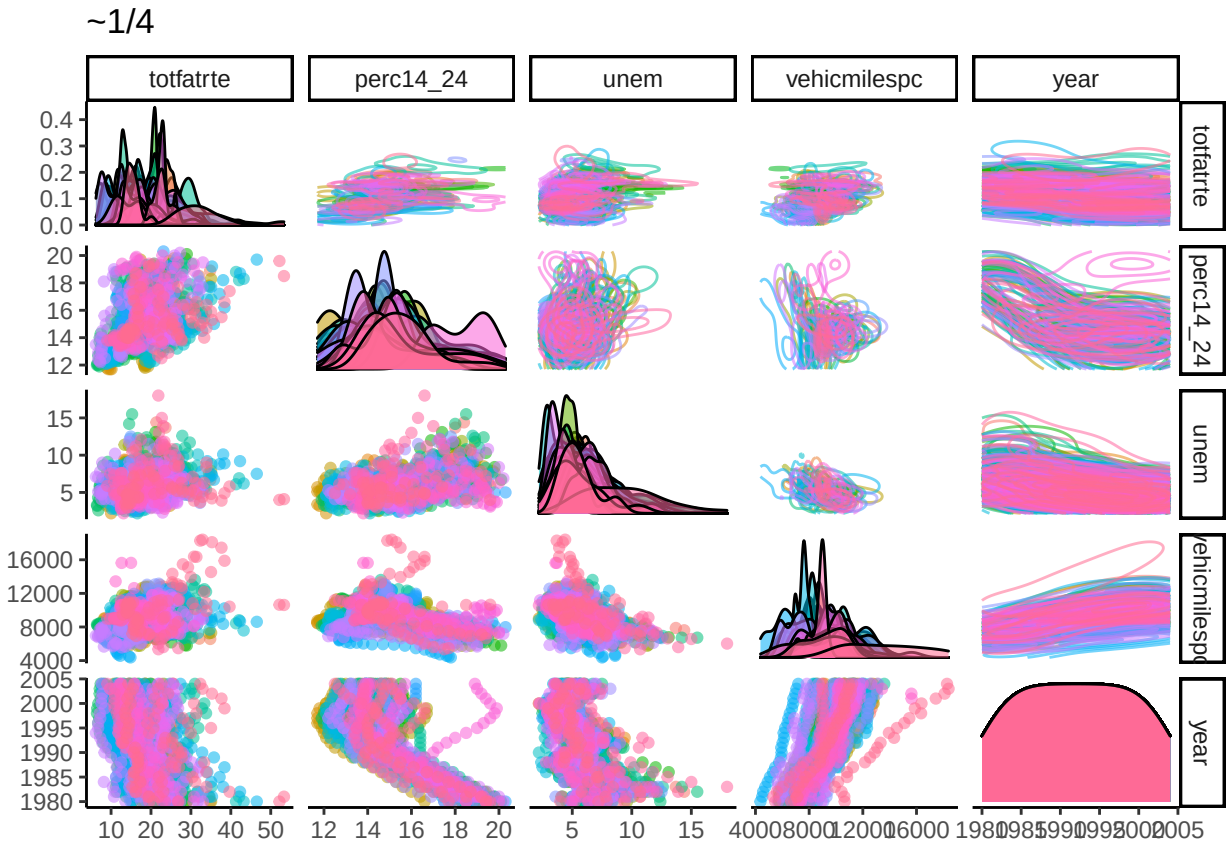


Figure 6: Relationship between the totfatrte and several potentially important predictors

Question 4

- Scatterplot matrix

```
# # Scatterplot matrix
# GGally::ggpairs(df,
#         aes( alpha=1/2),
#         mapping=ggplot2::aes(colour = factor(state)),
#         columns=c("totfatrte","d81","d82", "d83","d84", "d85","d86","d87","d88" , "d89" , "d
```

```
#          lower= list(continuous = "points"),
#          upper=list(continuous="blank"),
#          diag=list(continuous="blankDiag"))+
#          theme_classic()
```

- correlations of variables used in models:

```
# Variables
corrs <- cor(df[,c("totfatrte","d81","d82", "d83","d84", "d85","d86","d87","d88" , "d89" , "d90
#
#round(corrs,2)
```

Question 5

- Comparing demeaned fixed effects for each state

```
summary(fixef(FE.mod.4, type="dmean"))
```

```
##      Estimate Std. Error t-value Pr(>|t|)
## 1   0.2415138  0.4730017  0.5106   0.6097
## 3   0.2838662  0.4614069  0.6152   0.5385
## 4   0.2956770  0.4657175  0.6349   0.5256
## 5  -0.0804904  0.4613926 -0.1745   0.8615
## 6  -0.0451341  0.4621836 -0.0977   0.9222
## 7  -0.3335991  0.4567984 -0.7303   0.4654
## 8  -0.0414929  0.4673186 -0.0888   0.9293
## 10  0.2218976  0.4615416  0.4808   0.6308
## 11  0.0468065  0.4729032  0.0990   0.9212
## 13  0.2248554  0.4665102  0.4820   0.6299
## 14 -0.1191627  0.4541820 -0.2624   0.7931
## 15 -0.0607212  0.4682050 -0.1297   0.8968
## 16  0.0026656  0.4610418  0.0058   0.9954
## 17 -0.0198992  0.4635004 -0.0429   0.9658
## 18  0.1429434  0.4677607  0.3056   0.7600
## 19  0.2990063  0.4614881  0.6479   0.5172
## 20 -0.0657018  0.4659827 -0.1410   0.8879
## 21 -0.2022233  0.4606588 -0.4390   0.6608
## 22 -0.5865856  0.4549199 -1.2894   0.1975
## 23 -0.1143636  0.4632852 -0.2469   0.8051
## 24 -0.3113606  0.4632144 -0.6722   0.5016
## 25  0.4775844  0.4695268  1.0172   0.3093
## 26  0.0799385  0.4679776  0.1708   0.8644
## 27  0.3027704  0.4701088  0.6440   0.5197
## 28 -0.1182090  0.4621269 -0.2558   0.7982
## 29  0.3025459  0.4599275  0.6578   0.5108
```

```
## 30 -0.2887578  0.4616383 -0.6255   0.5318
## 31 -0.3829024  0.4548656 -0.8418   0.4001
## 32  0.4645654  0.4752394  0.9775   0.3285
## 33 -0.2428273  0.4429692 -0.5482   0.5837
## 34  0.1436164  0.4664837  0.3079   0.7582
## 35 -0.2451441  0.4664205 -0.5256   0.5993
## 36 -0.1678499  0.4603422 -0.3646   0.7155
## 37  0.0866716  0.4736028  0.1830   0.8548
## 38  0.1041295  0.4652681  0.2238   0.8229
## 39 -0.1104173  0.4537052 -0.2434   0.8078
## 40 -0.5359271  0.4497558 -1.1916   0.2337
## 41  0.2726773  0.4683129  0.5823   0.5605
## 42  0.0289322  0.4678246  0.0618   0.9507
## 43  0.1609962  0.4683363  0.3438   0.7311
## 44  0.0729323  0.4670759  0.1561   0.8759
## 45 -0.1268873  0.4625721 -0.2743   0.7839
## 46 -0.1273754  0.4698690 -0.2711   0.7864
## 47 -0.2499477  0.4655910 -0.5368   0.5915
## 48 -0.2052762  0.4624940 -0.4438   0.6572
## 49  0.3882397  0.4639389  0.8368   0.4029
## 50 -0.1602867  0.4646448 -0.3450   0.7302
## 51  0.2977111  0.4843519  0.6147   0.5389
```

```
summary(fixef(FE.no.time, type="dmean"))
```

```
##      Estimate Std. Error t-value Pr(>|t|)
## 1   0.2415138  0.4831229  0.4999   0.6172
## 3   0.2838662  0.4715647  0.6020   0.5473
## 4   0.2956770  0.4758635  0.6213   0.5345
## 5  -0.0804904  0.4715749 -0.1707   0.8645
## 6  -0.0451341  0.4724139 -0.0955   0.9239
## 7  -0.3335991  0.4670688 -0.7142   0.4752
## 8  -0.0414929  0.4774414 -0.0869   0.9308
## 10  0.2218976  0.4720609  0.4701   0.6384
## 11  0.0468065  0.4830651  0.0969   0.9228
## 13  0.2248554  0.4766024  0.4718   0.6372
## 14 -0.1191627  0.4642929 -0.2567   0.7975
## 15 -0.0607212  0.4783164 -0.1269   0.8990
## 16  0.0026656  0.4712310  0.0057   0.9955
## 17 -0.0198992  0.4737273 -0.0420   0.9665
## 18  0.1429434  0.4776709  0.2993   0.7648
## 19  0.2990063  0.4714074  0.6343   0.5260
## 20 -0.0657018  0.4762237 -0.1380   0.8903
## 21 -0.2022233  0.4708112 -0.4295   0.6676
## 22 -0.5865856  0.4648776 -1.2618   0.2073
## 23 -0.1143636  0.4733689 -0.2416   0.8091
## 24 -0.3113606  0.4734582 -0.6576   0.5109
```

```
## 25  0.4775844  0.4794645  0.9961   0.3194
## 26  0.0799385  0.4782784  0.1671   0.8673
## 27  0.3027704  0.4803373  0.6303   0.5286
## 28 -0.1182090  0.4723470 -0.2503   0.8024
## 29  0.3025459  0.4702918  0.6433   0.5202
## 30 -0.2887578  0.4717802 -0.6121   0.5406
## 31 -0.3829024  0.4650974 -0.8233   0.4105
## 32  0.4645654  0.4854267  0.9570   0.3388
## 33 -0.2428273  0.4530400 -0.5360   0.5921
## 34  0.1436164  0.4767234  0.3013   0.7633
## 35 -0.2451441  0.4764671 -0.5145   0.6070
## 36 -0.1678499  0.4704385 -0.3568   0.7213
## 37  0.0866716  0.4838596  0.1791   0.8579
## 38  0.1041295  0.4755784  0.2190   0.8267
## 39 -0.1104173  0.4638346 -0.2381   0.8119
## 40 -0.5359271  0.4597452 -1.1657   0.2440
## 41  0.2726773  0.4782538  0.5702   0.5687
## 42  0.0289322  0.4779633  0.0605   0.9517
## 43  0.1609962  0.4784222  0.3365   0.7365
## 44  0.0729323  0.4771593  0.1528   0.8785
## 45 -0.1268873  0.4723947 -0.2686   0.7883
## 46 -0.1273754  0.4800562 -0.2653   0.7908
## 47 -0.2499477  0.4757063 -0.5254   0.5994
## 48 -0.2052762  0.4727525 -0.4342   0.6642
## 49  0.3882397  0.4740031  0.8191   0.4129
## 50 -0.1602867  0.4747432 -0.3376   0.7357
## 51  0.2977111  0.4945771  0.6020   0.5473
```

Diagnostic plots for models

```
# 2
RvF.2 <- ggplot(data=df) +
  geom_point(aes(y=residuals(pOLS.mod.2), x=fitted(pOLS.mod.2))) +
  theme_classic() + stat_smooth(method="loess",aes(y=residuals(pOLS.mod.2), x=fitted(pOLS.mod.2
  geom_hline(yintercept=0, col="red", linetype="dashed") +
  labs(title="pooled OLS",x="fitted values",y="residuals")
QQ.2 <- ggplot(df, aes(sample=pOLS.mod.2$residuals)) +
  stat_qq() + stat_qq_line(color="blue") + theme_classic()
# 3
RvF.3 <- ggplot(data=df) +
  geom_point(aes(y=residuals(pOLS.mod.3), x=fitted(pOLS.mod.3))) +
  theme_classic() +stat_smooth(method="loess",aes(y=residuals(pOLS.mod.3), x=fitted(pOLS.mod.3
  geom_hline(yintercept=0, col="red", linetype="dashed")
  labs(title="expanded pooled OLS",x="fitted values",y="residuals")
```

```
## $x
```

```
## [1] "fitted values"
##
## $y
## [1] "residuals"
##
## $title
## [1] "expanded pooled OLS"
##
## attr(,"class")
## [1] "labels"
```

```
QQ.3 <- ggplot(df, aes(sample=pOLS.mod.3$residuals)) +
  stat_qq() + stat_qq_line(color="blue") + theme_classic()
# 4
RvF.4 <- ggplot(data=df) +
  geom_point(aes(y=residuals(FE.mod.4), x=fitted(FE.mod.4))) +
  theme_classic() +stat_smooth(method="loess",aes(y=residuals(FE.mod.4), x=fitted(FE.mod.4))) +
  geom_hline(yintercept=0, col="red", linetype="dashed")
  labs(title="fixed effects model",x="fitted values",y="residuals")
```

```
## $x
## [1] "fitted values"
##
## $y
## [1] "residuals"
##
## $title
## [1] "fixed effects model"
##
## attr(,"class")
## [1] "labels"
```

```
QQ.4 <- ggplot(df, aes(sample=FE.mod.4$residuals)) +
  stat_qq() + stat_qq_line(color="blue") + theme_classic()
# 5
RvF.5 <- ggplot(data=df) +
  geom_point(aes(y=residuals(RE.mod.5), x=fitted(RE.mod.5))) +
  theme_classic() +stat_smooth(method="loess",aes(y=residuals(RE.mod.5), x=fitted(RE.mod.5))) +
  geom_hline(yintercept=0, col="red", linetype="dashed")
  labs(title="random effects model",x="fitted values",y="residuals")
```

```
## $x
## [1] "fitted values"
##
## $y
## [1] "residuals"
```

```
##
## $title
## [1] "random effects model"
##
## attr(,"class")
## [1] "labels"
```

```
QQ.5 <- ggplot(df, aes(sample=RE.mod.5$residuals)) +
  stat_qq() + stat_qq_line(color="blue") + theme_classic()
# Arrange
grid.arrange(RvF.2,QQ.2,
             RvF.3,QQ.3,
             RvF.4,QQ.4,
             RvF.5,QQ.5,
             ncol=2)
```

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

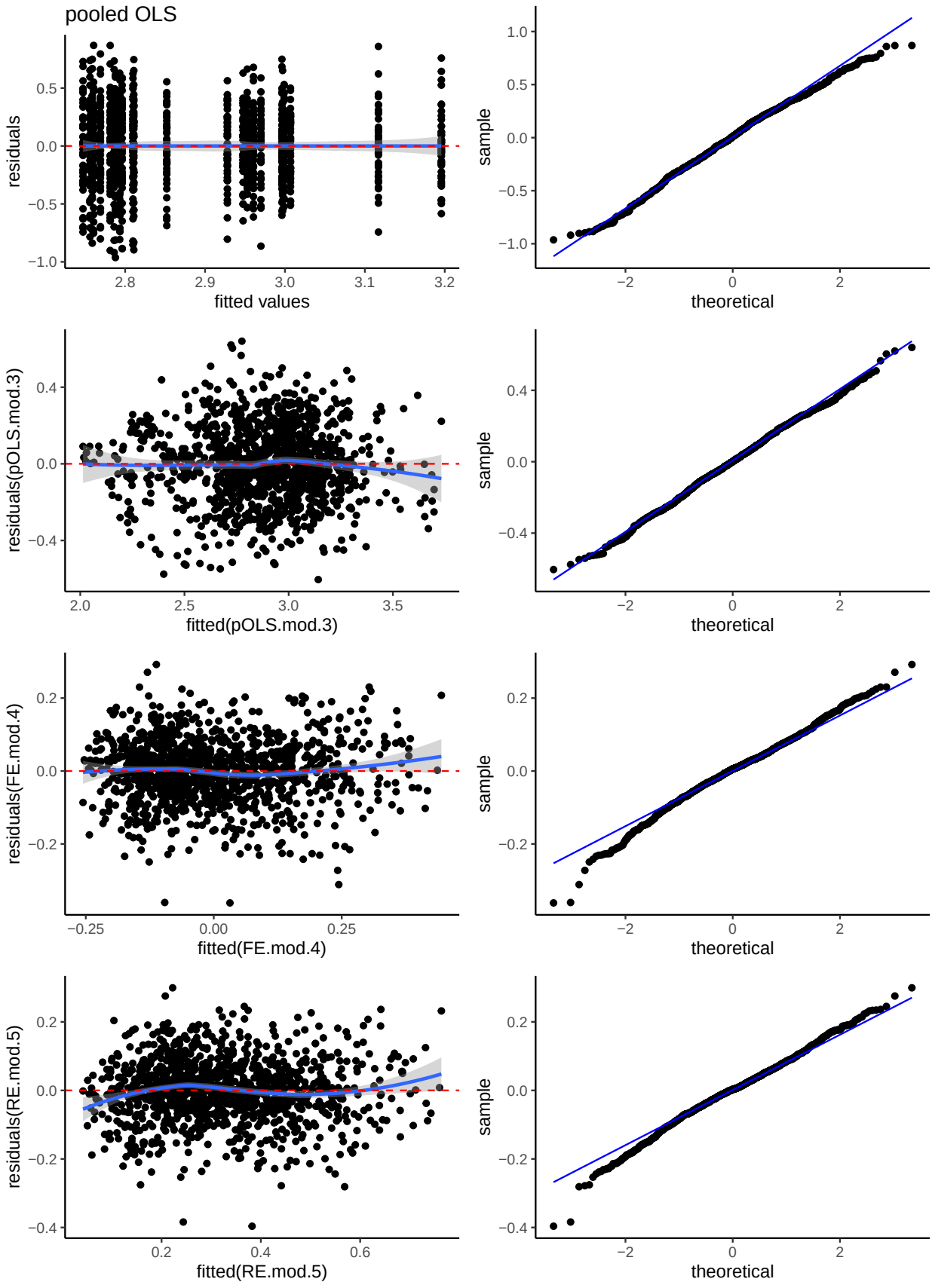


Figure 7: Diagnostic plots