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

Michael Berger, John Gao and Thomas Hamnett

Contents

Introduction **2**

Exercises **2**

Exploratory Data Analysis (EDA) 2
Response Variable and Linear Regression 11
Transformations and Linear Regression 13
Pooled and Fixed Effect Regression 16
Random or Fixed Effects? 19
Fixed Effects Prediction 20
Consequences of Heteroskedasticity in Idiosyncratic Errors 21

Conclusion **21**

Introduction

In this lab we investigate the question **"Do changes in traffic laws affect traffic fatalities?"** We are using the provided *driving.Rdata*, which includes 25 years of data for 48 continental U.S. states with different sets of laws including drunk driving limits, seat belt, and speed limit laws as well as other indicator, economic and demographic variables.

First, we will load the necessary libraries, the data and conduct a thorough Exploratory Data Analaysis (EDA). Next, we will clarify how the response variable *totfatrte* is defined and estimate a linear regression model on a set of dummy variables for the years. Then we expand the linear regression model including explanatory variables like per se law and primary and secondary seatbelt laws. After that we fit a pooled and fixed effect regression model, compare the results and discuss the assumptions. Then we compare random effects and fixed effects models. Finally, we predict the effect of an increase of *vehicmilespc* by 1,000 on *totfatrte* and discuss the impact of the presence of serial correlation of heteroskedasticity on the idiosyncratic error term on the different models' assumptions. We will use the results from the exercises to inform the question of interest: **"Do changes in traffic laws affect traffic fatalities?"**

Exercises

Exploratory Data Analysis (EDA)

Exercise: 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.

Answer: First, we load the libraries we use as well as the data.

```
library(knitr); opts_chunk$set(tidy.opts=list(width.cutoff=60),
                               tidy=T,warning=FALSE,message=FALSE)
```

```
rm(list = ls())
libs <- c("ggplot2", "ggfortify", "plotly", "dplyr", "astsa",
    "fpp2", "tidyr", "tseries", "forecast", "gridExtra", "plm",
    "lattice", "car", "Hmisc", "stargazer", "kableExtra", "reshape2",
    "pander", "xtable")
for (lib in libs) {
    require(lib, character.only = TRUE)
}
```

```
# Loading the data and inspecting structure
load("driving.RData")
df <- data
data.frame(variable = names(df), class = sapply(df, typeof),
    first_values = sapply(df, function(x) paste0(round(head(x),
        3), collapse = ", ")), row.names = NULL) %>% kable() %>%
    kable_styling(font_size = 9)
```

variable	class	first_values
year	integer	1980, 1981, 1982, 1983, 1984, 1985
state	integer	1, 1, 1, 1, 1, 1
sl55	double	1, 1, 1, 1, 1, 1
sl65	double	0, 0, 0, 0, 0, 0
sl70	double	0, 0, 0, 0, 0, 0
sl75	double	0, 0, 0, 0, 0, 0
slnone	double	0, 0, 0, 0, 0, 0
seatbelt	integer	0, 0, 0, 0, 0, 0
minage	double	18, 18, 18, 18, 18, 20
zerotol	double	0, 0, 0, 0, 0, 0
gdl	double	0, 0, 0, 0, 0, 0
bac10	double	1, 1, 1, 1, 1, 1
bac08	double	0, 0, 0, 0, 0, 0
perse	double	0, 0, 0, 0, 0, 0
totfat	integer	940, 933, 839, 930, 932, 882
nghtfat	integer	422, 434, 376, 397, 421, 358
wkndfat	integer	236, 248, 224, 223, 237, 224
totfatpvm	double	3.2, 3.35, 2.81, 3, 2.83, 2.51
nghtfatpvm	double	1.437, 1.558, 1.259, 1.281, 1.278, 1.019
wkndfatpvm	double	0.803, 0.89, 0.75, 0.719, 0.72, 0.637
statepop	integer	3893888, 3918520, 3925218, 3934109, 3951834, 3972527
totfatrte	double	24.14, 24.07, 21.37, 23.64, 23.58, 22.2
nghtfatrte	double	10.84, 11.08, 9.58, 10.09, 10.65, 9.01
wkndfatrte	double	6.06, 6.33, 5.71, 5.67, 6, 5.64
vehicmiles	double	29.375, 27.852, 29.858, 31, 32.933, 35.139
unem	double	8.8, 10.7, 14.4, 13.7, 11.1, 8.9
perc14_24	double	18.9, 18.7, 18.4, 18, 17.6, 17.3
sl70plus	double	0, 0, 0, 0, 0, 0
sbprim	integer	0, 0, 0, 0, 0, 0
sbsecon	integer	0, 0, 0, 0, 0, 0
d80	integer	1, 0, 0, 0, 0, 0
d81	integer	0, 1, 0, 0, 0, 0
d82	integer	0, 0, 1, 0, 0, 0
d83	integer	0, 0, 0, 1, 0, 0
d84	integer	0, 0, 0, 0, 1, 0
d85	integer	0, 0, 0, 0, 0, 1
d86	integer	0, 0, 0, 0, 0, 0
d87	integer	0, 0, 0, 0, 0, 0
d88	integer	0, 0, 0, 0, 0, 0
d89	integer	0, 0, 0, 0, 0, 0
d90	integer	0, 0, 0, 0, 0, 0
d91	integer	0, 0, 0, 0, 0, 0
d92	integer	0, 0, 0, 0, 0, 0
d93	integer	0, 0, 0, 0, 0, 0
d94	integer	0, 0, 0, 0, 0, 0
d95	integer	0, 0, 0, 0, 0, 0
d96	integer	0, 0, 0, 0, 0, 0
d97	integer	0, 0, 0, 0, 0, 0
d98	integer	0, 0, 0, 0, 0, 0
d99	integer	0, 0, 0, 0, 0, 0
d00	integer	0, 0, 0, 0, 0, 0
d01	integer	0, 0, 0, 0, 0, 0
d02	integer	0, 0, 0, 0, 0, 0
d03	integer	0, 0, 0, 0, 0, 0
d04	integer	0, 0, 0, 0, 0, 0
vehicmilespc	double	7543.874, 7107.785, 7606.622, 7879.802, 8333.562, 8845.614

```
# summary statistics of some variables
kable(summary(df[, c("totfatrte", "statepop", "unem", "sl70plus",
    "seatbelt", "vehicmilespc", "perc14_24")]), "latex", booktabs = T) %>%
    kable_styling(font_size = 8)
```

totfatrte	statepop	unem	sl70plus	seatbelt	vehicmilespc	perc14_24
Min. : 6.20	Min. : 453401	Min. : 2.200	Min. :0.0000	Min. :0.000	Min. : 4372	Min. :11.70
1st Qu.:14.38	1st Qu.: 1641938	1st Qu.: 4.500	1st Qu.:0.0000	1st Qu.:0.000	1st Qu.: 7788	1st Qu.:13.90
Median :18.43	Median : 3700425	Median : 5.600	Median :0.0000	Median :1.000	Median : 9013	Median :14.90
Mean :18.92	Mean : 5329896	Mean : 5.951	Mean :0.2068	Mean :1.116	Mean : 9129	Mean :15.33
3rd Qu.:22.77	3rd Qu.: 6069563	3rd Qu.: 7.000	3rd Qu.:0.0000	3rd Qu.:2.000	3rd Qu.:10327	3rd Qu.:16.60
Max. :53.32	Max. :35894000	Max. :18.000	Max. :1.0000	Max. :2.000	Max. :18390	Max. :20.30

```
any(is.na(df))  # Checking for missing values
```

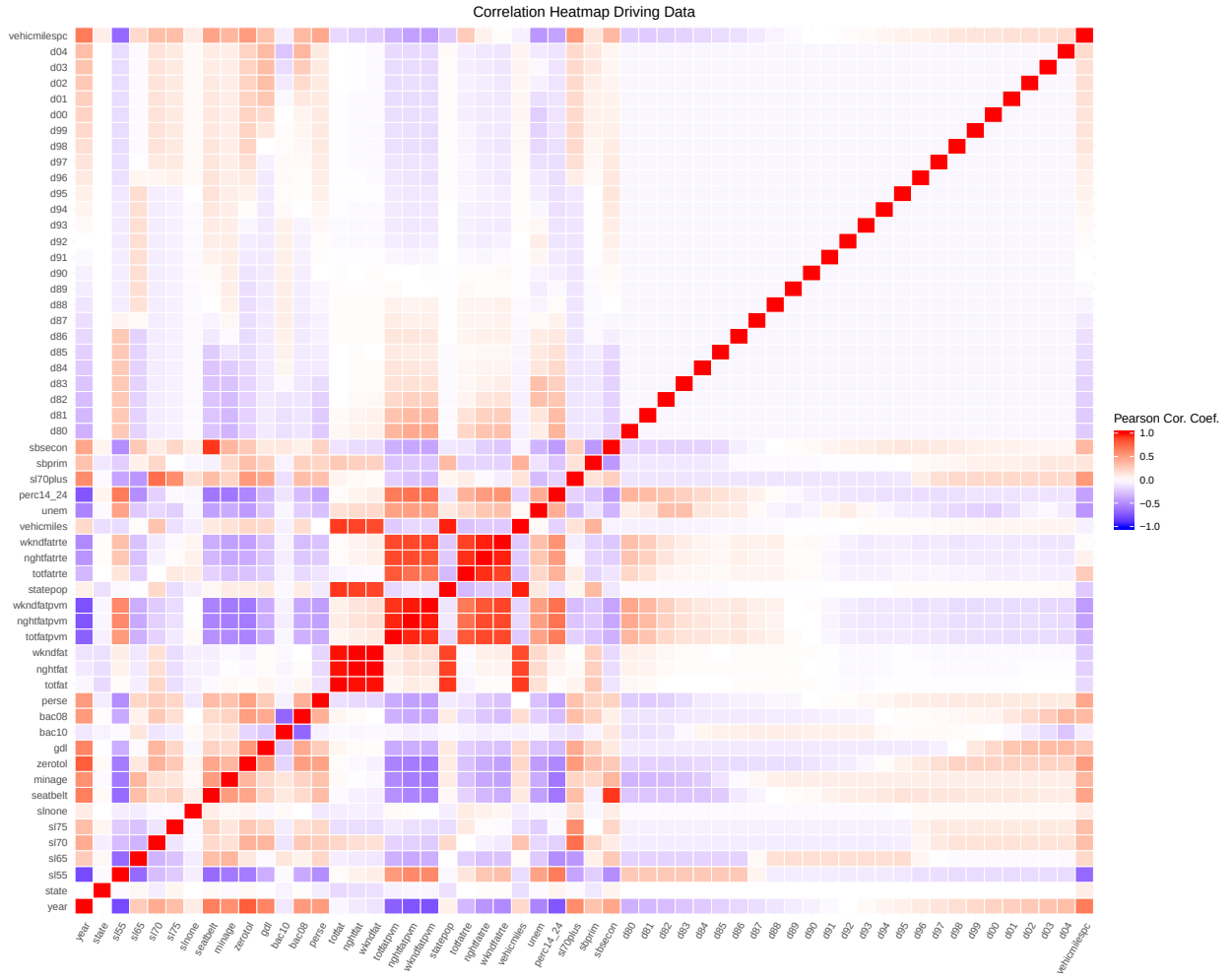
```
## [1] FALSE
```

we have 1,200 observations for 56 variables. All variables are coded as integers. We also see that we have dummy variables for the single years as well as a variable *year*. In addition, we have no missing values in the data. As we have so many variables, we display only the summary statistics of some which we consider a priori to be important in the further analysis and as representations for law, demographic and economic variables.

The response variable *totfatrte* has a minimum of 6.20 and a maximum of 53.32 with a median of 27.50 and a mean of 18.92. The states have different populations ranging from 453,401 to 35,894,000. The speed limit of 70+ and the presence of a seatbelt law are coded as categorical variables (seatbelt with three levels, corresponding to no law, primary and secondary) with the mean being 0.2068 and 1.116 respectively. The unemployment variable ranges from 2.2 to 18 and the vehicle miles per capita range from 4,372 to 18,390 with a mean of 9,129. Finally, the percentage of the population between 14 to 24 ranges from 11.7 to 20.3 with a mean of 15.33. Hence, we see some variation in the variables.

Next, we look at the correlations. As the variables are so many, we use a correlation heatmap to display the information.

```
# correlations heatmap
heatmap_data <- df

cormat <- round(cor(heatmap_data, use = "complete.obs"), 2)
m_cormat <- melt(cormat)
colnames(m_cormat) <- c("Variables_1", "Variables_2", "Values")

ggplot(data = m_cormat, aes(x = Variables_1, y = Variables_2,
    fill = Values)) + geom_tile(color = "white") + scale_fill_gradient2(low = "blue",
    high = "red", mid = "white", midpoint = 0, limit = c(-1,
        1), space = "Lab", name = "Pearson Cor. Coef.") + theme_minimal() +
    theme(axis.title.x = element_blank(), axis.title.y = element_blank(),
        axis.text.x = element_text(angle = 60, hjust = 1), plot.title = element_text(hjust = 0
    ggtitle("Correlation Heatmap Driving Data")
```



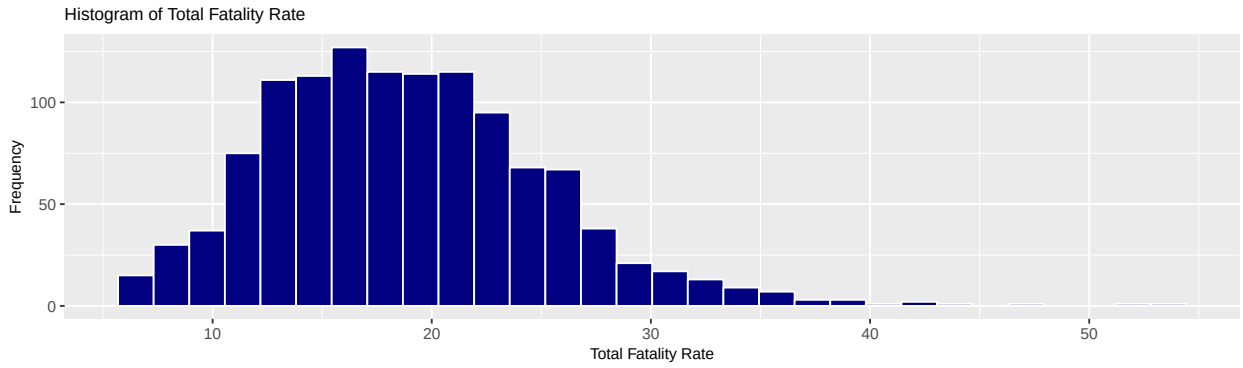
From the correlation heatmap, we see that *totfatrte* is positively correlated with vehicle miles per capita, shows different correlations with different states, is negatively correlated with year and negatively correlated with the presence of a seatbelt law, minimum drinking age, zero tolerance law and per se laws. It is also slightly negatively correlated with the state population. In addition, the variable is perfectly correlated with other variables measuring fatalities on different basis. These variables will be removed from the dataset.

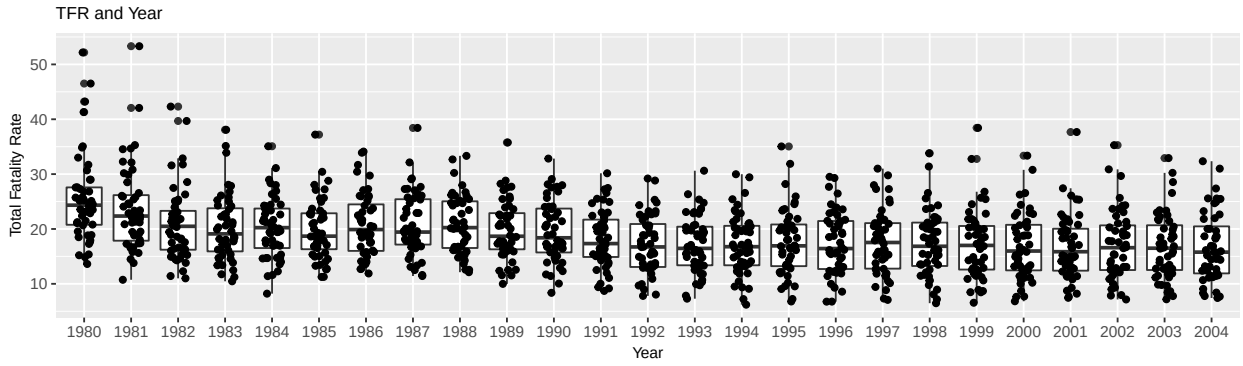
```
df_adj <- df[, -which(names(df) %in% c("totfat", "nghtfat", "wkndfat",
    "totfatpvm", "nghtfatpvm", "wkndfatpvm", "nghtfatrte", "wkndfatrte"))]
```

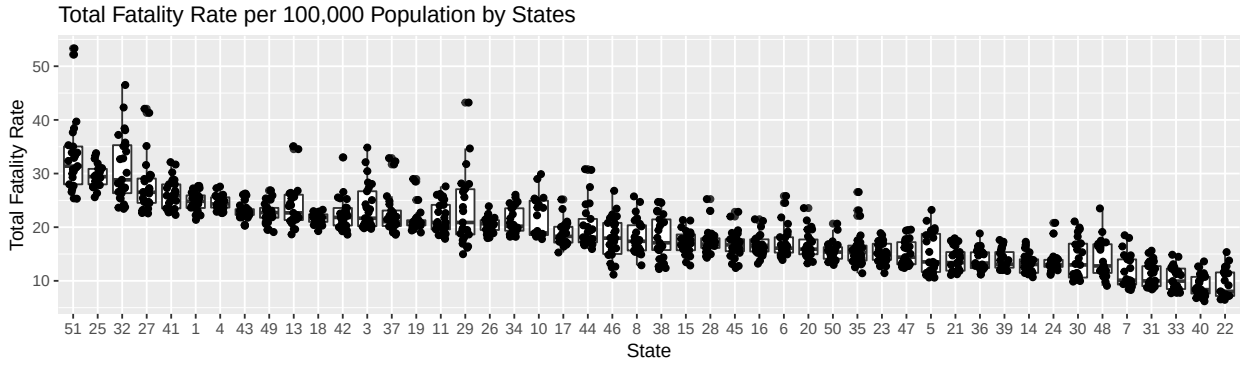
Next, we look at visualizations of *totfatrte* over time and state.

```
# helper functions
gBox <- function(var1, var2, input = df_adj, col2 = "totfatrte") {
    ggplot(input, aes(input[[var1]], input[[col2]])) + ggtitle(paste("TFR and",
        var2)) + ylab("Total Fatality Rate") + xlab(var2) + geom_boxplot() +
        geom_jitter(width = 0.2) + theme(plot.title = element_text(size = 10),
        axis.title = element_text(size = rel(0.8)))
}
gHist <- function(input = df_adj, var, xlbl) {
```

```
    ggplot(input, aes(x = input[[var]])) + ggtitle(paste("Histogram of",
        xlbl)) + geom_histogram(col = "white", fill = "navy",
        bins = 30) + xlab(xlbl) + ylab("Frequency") + theme(plot.title = element_text(size = 1
        axis.title = element_text(size = rel(0.8)))
}
# Death rate distribution, by Year, and by State
plot1 <- gHist(var = "totfatrte", xlbl = "Total Fatality Rate")
df_adj <- mutate(df_adj, yr_fac = as.factor(year))
plot2 <- gBox(var1 = "yr_fac", var2 = "Year")
plot3 <- ggplot(df_adj, aes(reorder(as.factor(state), -totfatrte,
    median), totfatrte)) + geom_boxplot() + geom_jitter(width = 0.2) +
    ylab("Total Fatality Rate") + xlab("State") + ggtitle("Total Fatality Rate per 100,000 Popu
grid.arrange(plot1, plot2, plot3, nrow = 3)
```







We see that *totfatrte* is slightly skewed to the left with a right tail. In addition, the boxplots over the years show some downward trend in the median of *totfatrte* as well as some change in the spread of

totfatrte. In addition, the boxplots over the states show some strong differences in *totfatrte* between states. This indicates that both states and time provide information about the variation of *totfatrte*. Hence, we cannot assume independent and identically distributed observations. Both, state and time dimension, should be considered in a model.

Next, we look at some explanatory variables and their association with *totfatrte*. We chose the variables based on the correlation heatmap. The below analysis related to the visualizations of the twelve explanatory variables in the below plots.

From below visual analysis of the twelve explanatory variables chosen, we see that the presence of a speed limit of 55 appears not to have made much of a difference in terms of median *totfatrte* values for the states where the speed limit was always present or absent. For the remaining cases, there is a slight trend towards a higher presence being associated with a lower *totfatrte*, but the observations are few for these cases.

The presence of seatbelt laws is associated with a lower *totfatrte*, whereas a level of 1 (primary) shows a lower median than a level of 2 (secondary). The observations seem also to be roughly equally distributed accross the levels 0 and 2. For level 1, only roughly half of the observations of either 0 or 2 are present.

From the visualizations we further note that a lower minimum drinking age does not indicate a strong association with lower *totfatrte*, besides the ages 18 and 21. The reason might be that most of the observations fall to 21.

Furthermore, the boxplots show that lower speed limits are associated with lower *totfatrte*, while a lower blood alcohol limit of 0.08 seems not necessarily be associated with a lower *totfatrte*. Graduated driving law being present over the whole time, a zero tolerance law as well as a per se law being present the whole time seem all to be associated with a lower *totfatrte*.

A higher percentage of the population being between 14 and 24 seems to be associated with a higher *totfatrte*, regardless of the presence of a graduate drivers law, which is only associated with a lower percentage of the population being in this age range.

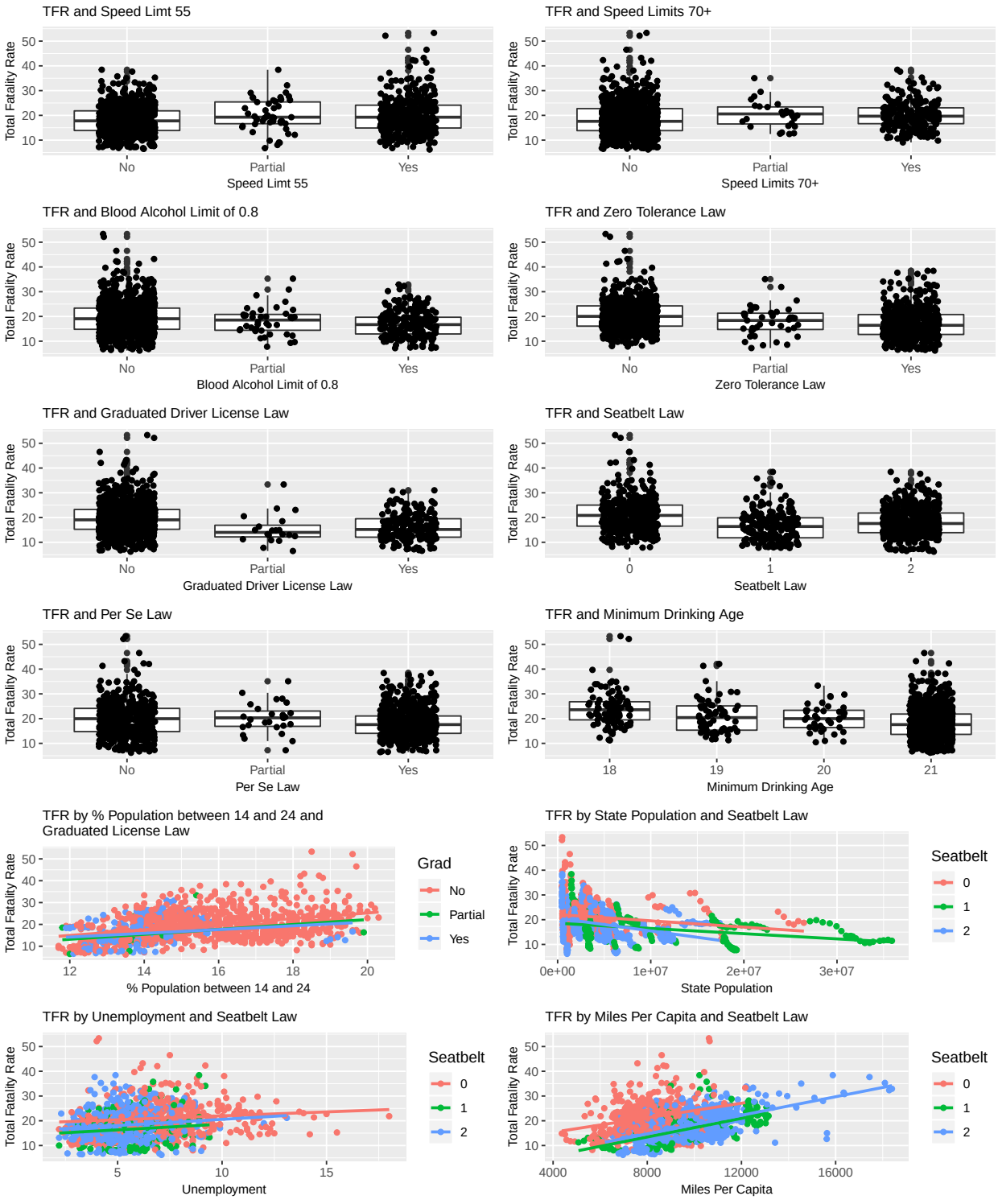
In addition, higher population states have less variability in *totfatrte* than lower population states, which we would expect even if there was no true difference in the population mean between the states (rate is more variable with smaller population). There is no clear evidence from the chart that the population-*totfatrte* relationship is influenced by seatbelt laws.

Higher unemployment rate seems to show a slight association with a higher *totfatrte*, some of the highest unemployment rates seem to be associated with no seatbelt laws. Finally, a higher vehicle miles per capita is associated both with a higher *totfatrte* as well as with seatbelt laws being present.

```
# Tabular analysis for seatbelt
kable(table(round(df_adj$seatbelt, 1)), "latex", col.names = c("seatbelt",
    "frequency"), booktabs = T) %>% kable_styling(font_size = 8)
```

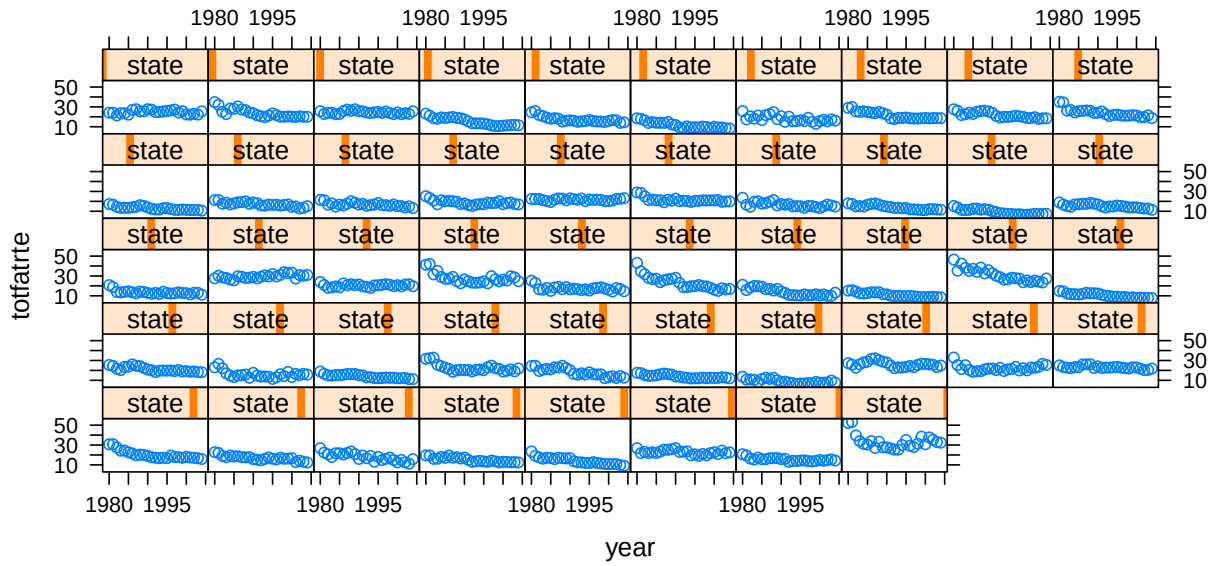
seatbelt	frequency
0	423
1	215
2	562

```r
# Binning Variables
dfbin <- function(df, nmvec) {
    for (i in 1:length(nmvec)) {
        vname <- paste0(nmvec[i], "bin")
        df[vname] <- cut(df[[nmvec[i]]], breaks = c(-1, 0, 0.999,
            1), labels = c("No", "Partial", "Yes"))
    }
    df
}
nms <- c("sl55", "sl70plus", "bac08", "zerotol", "gdl", "perse")
df_adj <- dfbin(df_adj, nms) %>% mutate(sbelt = as.factor(seatbelt))
# For partial years (change in law) for drinking age, round
# to nearest full year
df_adj["minagebin"] <- cut(df_adj$minage, breaks = c(-1, 18.5,
    19.5, 20.5, 21), labels = c("18", "19", "20", "21"))

bxs <- c("sl55bin", "sl70plusbin", "bac08bin", "zerotolbin",
    "gdlbin", "sbelt", "persebin", "minagebin")
xs <- c("Speed Limt 55", "Speed Limits 70+", "Blood Alcohol Limit of 0.8",
    "Zero Tolerance Law", "Graduated Driver License Law", "Seatbelt Law",
    "Per Se Law", "Minimum Drinking Age")
bxplts <- mapply(gBox, var1 = bxs, var2 = xs, SIMPLIFY = FALSE,
    USE.NAMES = FALSE)

gScat <- function(var1, var2, var3, var5, input = df_adj, var4 = "totfatrte") {
    ggplot(input, aes(input[[var1]], input[[var4]], colour = input[[var3]])) +
        geom_point() + geom_smooth(method = "lm", formula = y ~
        x, se = FALSE) + xlab(var2) + ggtitle(paste("TFR by",
        var2, var5)) + ylab("Total Fatality Rate") + scale_color_discrete(name = ifelse(var3 ==
        "gdlbin", "Grad", "Seatbelt")) + theme(plot.title = element_text(size = 10),
        axis.title = element_text(size = rel(0.8)))
}

clrs <- c("gdlbin", "sbelt", "sbelt", "sbelt")
sctrs <- c("perc14_24", "statepop", "unem", "vehicmilespc")
x2s <- c("% Population between 14 and 24", "State Population",
    "Unemployment", "Miles Per Capita")
ttls <- c("and \nGraduated License Law", "and Seatbelt Law",
    "and Seatbelt Law", "and Seatbelt Law")
sctplts <- mapply(gScat, var1 = sctrs, var2 = x2s, var3 = clrs,
    var5 = ttls, SIMPLIFY = FALSE, USE.NAMES = FALSE)
grid.arrange(grobs = c(bxplts, sctplts), nrow = 6)
```

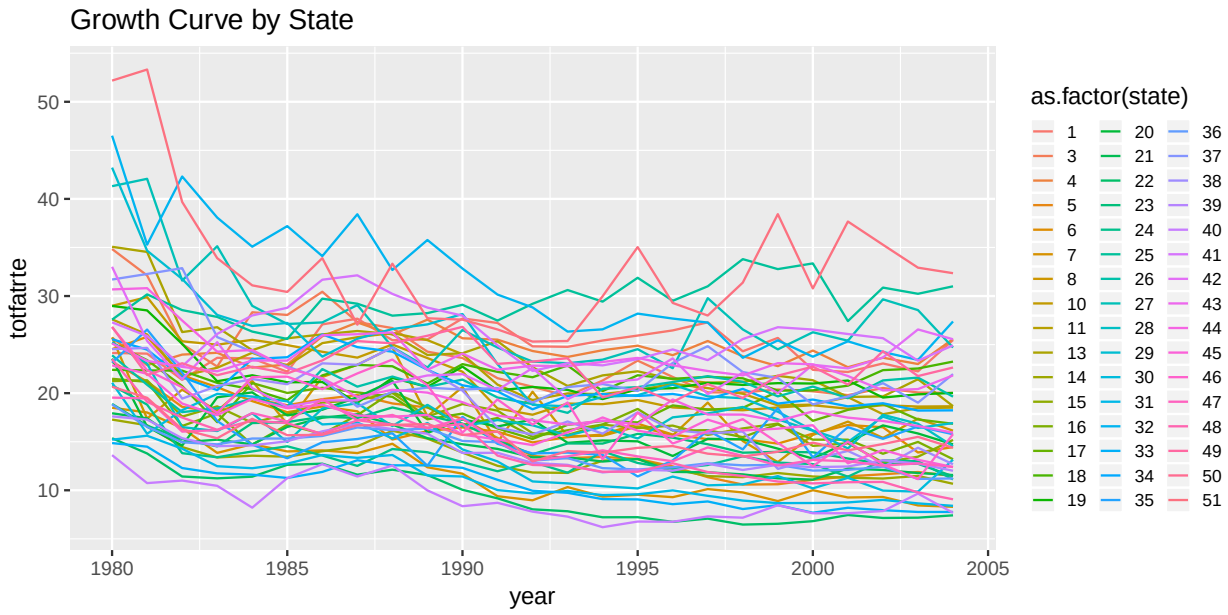


Next, we look at the growth curve to inform within-state trends across time.

```
# Growth Curve
g1 <- xyplot(totfatrte ~ year | state, data = df_adj, as.table = T)
g1
```




```
g2 <- ggplot(df_adj, aes(year, totfatrte, colour = as.factor(state))) +
    geom_line() + ggtitle("Growth Curve by State") + theme(legend.text = element_text(size = 8)
    legend.key.size = unit(0.4, "cm"))
g2
```



Both graphs indicate a downward trend accross almost all states in terms of *totfatrte*. In addition, different states indicate different *totfatrte* levels. Both plots again stress the importance to consider states as id and year as time variable to model *totfatrte*. In addition, it becomes again clear that the assumption about independence and identically distributed observations as well as the assumption about homoscedasticity are most likely violated. Thus, we will need to choose models which consider both factors and treat the panel data we have appropriately.

To summarize the findings of our EDA:

- The outcome variable *totfatrte* varies across time and states. This indicates that we are dealing with panel data and should choose models suitable to incorporate this.
- Some of the non-state and non-time explanatory variables indicate positive or negative correlations with *totfatrte*. In addition, there seem to exists associations between some of the explanatory variables. Some explanatory variables are law related, like *seatbelt*, and hence are of interest in the context of our question, while others, like *vehiclemilespc*, might serve as control variables.
- The growth curves indicate different *totfatrte* over states and time with, as it seems, a common downward trend.

Response Variable and Linear Regression

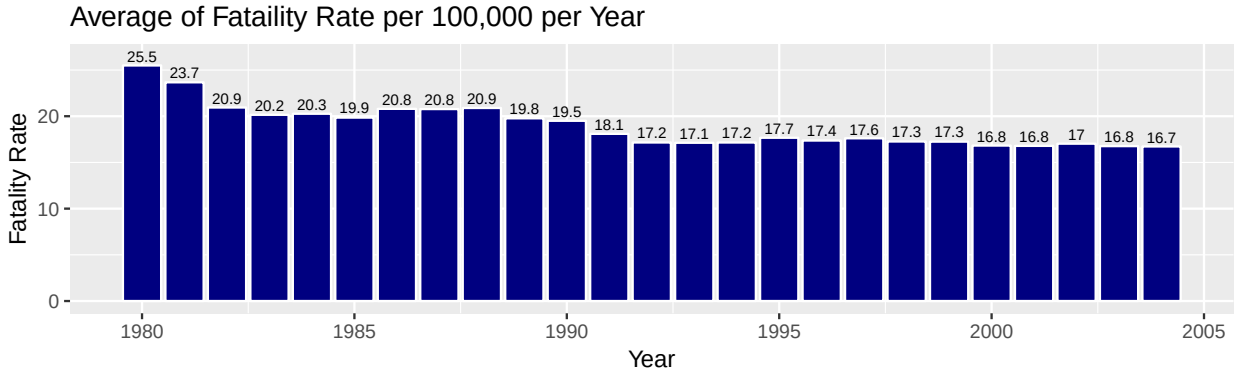
Exercise: How is our dependent variable of interest *totfatrte* defined? What is the average of this variable in each of the years in the time period covered in this dataset? Estimate a linear regression model of *totfatrte* on a set of dummy variables for the years 1981 through 2004. What does this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

Answer:

Looking at the variable description by calling `desc`, we see that *totfatrte* is defined as the total fatalities per 100,000 population. Hence, the variable is population adjusted.

We now look at the average of the variable at each of the years in the dataset.

```
ggplot(aggregate(df_adj$totfatrte, list(df_adj$year), mean),
    aes(Group.1, x, label = round(x, 1))) + xlab("Year") + ylab("Fatality Rate") +
    ggtitle("Average of Fatality Rate per 100,000 per Year") +
    geom_bar(stat = "identity", fill = "navy", colour = "white") +
    geom_text(size = 2.5, aes(label = round(x, 1), y = round(x,
        1) + 1))
```



We see that the average of the fatality rate is declining over the years, with a strong downswing in the beginning and at the end of the 1980s / beginning 1990s.

We now estimate a linear regression model of *totfatrte* on a set of dummy variables for the years 1981 through 2004. The model is described as follows:

$$\hat{totfatrte} = \beta_0 + \sum_{i=1981}^{2004} time_i \beta_i$$

This model explains the impact of each year on explaining the fatality rate.

```
model.lm <- lm(totfatrte ~ as.factor(year), data = df_adj)
pander(summary(model.lm))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	25.49	0.8671	29.4	7.267e-143
as.factor(year)1981	-1.824	1.226	-1.488	0.1371
as.factor(year)1982	-4.552	1.226	-3.712	0.0002152
as.factor(year)1983	-5.342	1.226	-4.356	1.44e-05
as.factor(year)1984	-5.227	1.226	-4.263	2.183e-05
as.factor(year)1985	-5.643	1.226	-4.602	4.644e-06
as.factor(year)1986	-4.694	1.226	-3.828	0.000136
as.factor(year)1987	-4.72	1.226	-3.849	0.0001251
as.factor(year)1988	-4.603	1.226	-3.754	0.0001829
as.factor(year)1989	-5.722	1.226	-4.666	3.418e-06
as.factor(year)1990	-5.989	1.226	-4.884	1.182e-06
as.factor(year)1991	-7.4	1.226	-6.034	2.137e-09
as.factor(year)1992	-8.337	1.226	-6.798	1.681e-11
as.factor(year)1993	-8.367	1.226	-6.823	1.425e-11
as.factor(year)1994	-8.339	1.226	-6.8	1.656e-11
as.factor(year)1995	-7.826	1.226	-6.382	2.512e-10
as.factor(year)1996	-8.125	1.226	-6.626	5.246e-11
as.factor(year)1997	-7.884	1.226	-6.429	1.863e-10
as.factor(year)1998	-8.229	1.226	-6.711	3.007e-11
as.factor(year)1999	-8.244	1.226	-6.723	2.774e-11
as.factor(year)2000	-8.669	1.226	-7.069	2.666e-12
as.factor(year)2001	-8.702	1.226	-7.096	2.214e-12
as.factor(year)2002	-8.465	1.226	-6.903	8.316e-12
as.factor(year)2003	-8.731	1.226	-7.12	1.877e-12
as.factor(year)2004	-8.766	1.226	-7.148	1.542e-12

Table 2: Fitting linear model: totfatrte ~ as.factor(year)

Observations	Residual Std. Error	R^2	Adjusted R^2
1200	6.008	0.1276	0.1098

We see from the model estimates that driving seems to become safer with each year, as the coefficients are negative. They also become more negative with each year relative to the **baseline year 1980**. Besides 1981, each year shows a high significance level. However, due to serial correlation and the violation of the iid assumption, the statistics are invalid. Hence, we cannot make a claim

of causality.

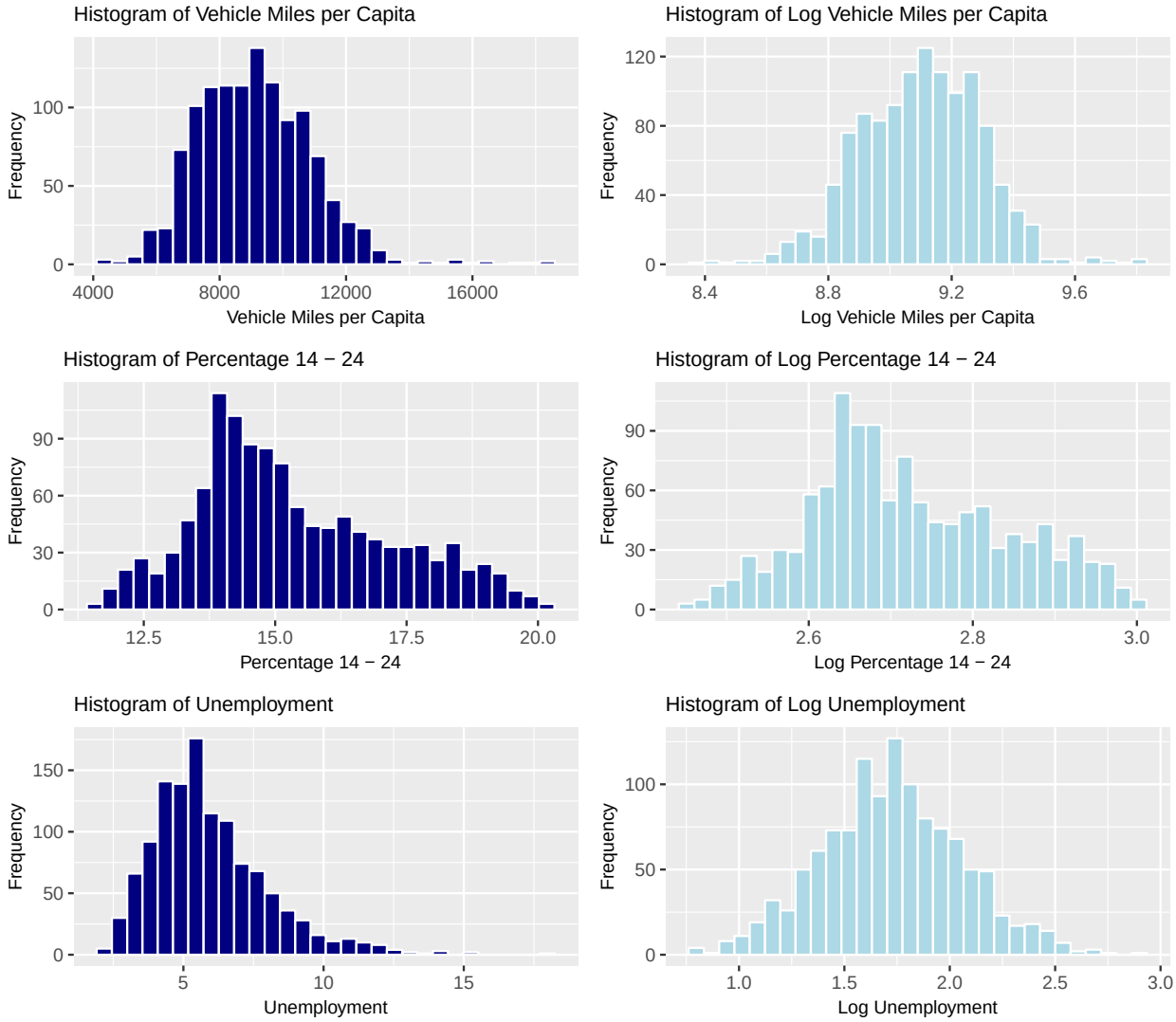
Transformations and Linear Regression

Exercise: 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.)

Answer:

The variables *bac08*, *bac10*, *perse*, *sbprim*, *sbsecon*, *sl70plus*, and *gdl* should in our mind not be transformed as they represent the percentage of the presence of blood alcohol level 0.08, 0.10, per se law, primary seatbelt law, secondary seatbelt law, speed limit of 70 plus and graduated drivers law in the data rather than dummy variables. The variables *vehicmilespc*, *perc14_24* and *unem* might be log-transformed to capture outliers better and express the three variables as changes in base levels for interpretation. As can be seen by the histogram plots below, especially *unem* looks more normally distributed after the transformation and the skewness of *vehicmilespc* and *perc14_24* is slightly reduced. The time series plots did not change much, besides indicating that outliers are captured a bit better. The log-transformation just led to a rescaling, but did not change the shape of the series. Hence, we did not include the time series plots in this report.

```
gHistLog <- function(input = df_adj, var, xlbl) {
    ggplot(input, aes(x = log(input[[var]]))) + ggtitle(paste("Histogram of Log",
        xlbl)) + geom_histogram(col = "white", fill = "lightblue",
        bins = 30) + xlab(paste("Log", xlbl)) + ylab("Frequency") +
        theme(plot.title = element_text(size = 10), axis.title = element_text(size = rel(0.8)))
}
vars <- c("vehicmilespc", "perc14_24", "unem")
xlbls <- c("Vehicle Miles per Capita", "Percentage 14 - 24",
    "Unemployment")
histplts <- mapply(gHist, var = vars, xlbl = xlbls, SIMPLIFY = FALSE,
    USE.NAMES = FALSE)
histlgplts <- mapply(gHistLog, var = vars, xlbl = xlbls, SIMPLIFY = FALSE,
    USE.NAMES = FALSE)
lay <- rbind(c(1, 4), c(2, 5), c(3, 6))
grid.arrange(grobs = c(histplts, histlgplts), layout_matrix = lay)
```




```
lg_trans <- c("vehicmilespc", "perc14_24", "unem")
for (c in lg_trans) {
    df_adj[, paste0(c, "log")] <- log(df_adj[, c])
}
```

We now estimate the model. The model becomes:

$$\hat{totfatrte} = \beta_0 + \sum_{i=1981}^{2004} time_i\beta_i + \sum var_j\beta_j$$

with each var_j being one of the transformed additional variables *bac08, bac10, perse, sbprim, sbsecon, sl70plus, gdl, perc14_24log, unemlog, vehicmilespclog.*

```
model.lm2 <- lm(totfatrte ~ as.factor(year) + bac08 + bac10 +
    perse + sbprim + sbsecon + sl70plus + gdl + perc14_24log +
    unemlog + vehicmilespclog, data = df_adj)
```

```
pander(summary(model.lm2))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-242.3	8.71	-27.82	4.905e-131
as.factor(year)1981	-2.144	0.8043	-2.666	0.007778
as.factor(year)1982	-6.503	0.8199	-7.931	5.081e-15
as.factor(year)1983	-7.448	0.8331	-8.94	1.495e-18
as.factor(year)1984	-6.262	0.8469	-7.394	2.723e-13
as.factor(year)1985	-7.028	0.8633	-8.141	9.972e-16
as.factor(year)1986	-6.409	0.8977	-7.139	1.652e-12
as.factor(year)1987	-7.017	0.9338	-7.514	1.138e-13
as.factor(year)1988	-7.215	0.9809	-7.356	3.576e-13
as.factor(year)1989	-8.812	1.018	-8.655	1.618e-17
as.factor(year)1990	-9.842	1.041	-9.457	1.689e-20
as.factor(year)1991	-12.03	1.064	-11.31	3.261e-28
as.factor(year)1992	-13.84	1.085	-12.75	5.933e-35
as.factor(year)1993	-13.64	1.1	-12.4	3.021e-33
as.factor(year)1994	-13.15	1.122	-11.72	4.727e-30
as.factor(year)1995	-12.58	1.15	-10.94	1.428e-26
as.factor(year)1996	-14.47	1.192	-12.14	4.947e-32
as.factor(year)1997	-14.6	1.221	-11.96	3.592e-31
as.factor(year)1998	-15.16	1.239	-12.23	1.866e-32
as.factor(year)1999	-14.97	1.259	-11.89	7.222e-31
as.factor(year)2000	-15.12	1.28	-11.81	1.719e-30
as.factor(year)2001	-16.05	1.297	-12.37	3.917e-33
as.factor(year)2002	-16.78	1.304	-12.86	1.697e-35
as.factor(year)2003	-17.14	1.315	-13.04	2.189e-36
as.factor(year)2004	-16.6	1.343	-12.37	4.34e-33
bac08	-2.55	0.5228	-4.878	1.218e-06
bac10	-1.31	0.3853	-3.4	0.0006962
perse	-0.6634	0.2901	-2.287	0.02238
sbprim	-0.07241	0.4803	-0.1508	0.8802
sbsecon	0.02595	0.4188	0.06197	0.9506
sl70plus	3.322	0.4345	7.645	4.361e-14
gdl	-0.8859	0.5108	-1.735	0.08309
perc14_24log	2.988	1.817	1.645	0.1003
unemlog	5.347	0.4718	11.33	2.537e-28
vehicmilespclog	28.14	0.8716	32.29	7.209e-164

Table 4: Fitting linear model: totfatrte ~ as.factor(year) + bac08 + bac10 + perse + sbprim + sbsecon + sl70plus + gdl + perc14_24log + unemlog + vehicmilespclog

Observations	Residual Std. Error	R^2	Adjusted R^2
1200	3.935	0.629	0.6182

The two variables for blood alcohol limits, *bac10* and *bac08*, are defined as blood alcohol limit .10 and blood alcohol limit .08, respectively.

We see from the regression results that the coefficients for the years become more negative. The logs of *vehicmilespc*, *perc14_24* and *unem* as well as the secondary seatbelt law are positively associated with fatality rates. Hence, not all laws have a negative effect. Coming back to our original question, whether changes in traffic laws affect traffic fatalities, we can say that based on this model changes in traffic laws in itself do not affect traffic fatalities. It depends on which traffic laws are changed.

bac08 has a more negative coefficient than *bac10* indicating that a stricter blood alcohol limit is associated with a lower fatality rate. The presence of per se law and primary seatbelt laws are also associated with a lower fatality rate.

The significance levels can again not be interpreted due to the violation of the OLS assumptions, as mentioned in the previous question.

Pooled and Fixed Effect Regression

Exercise: 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?

Answer:

We now repeat the above exercise with a fixed effect and pooled regression. For the fixed effect model, the model is:

$$\hat{totfatrte} - \overline{totfatrte} = \sum \beta_i (var_i - \overline{var_i})$$

with each var_i being one of the transformed variables *year, bac08, bac10, perse, sbprim, sbsecon, sl70plus, gdl, perc14_24log, unemlog, vehicmilespclog.* The index variables are *state* and *year.*

Hence, we estimate the effect of changes in the variables on the change of the total fatality rate. In addition, constant or nearly constant variables over time will be differenced out.

In the pool regression we estimate the panels by using a composite error form, assuming the time invariant omitted variables are not correlated with our explanatory variables and no serial correlation is present.

```
df.panel <- pdata.frame(df_adj, c("state", "year"))

model.pooled <- plm(totfatrte ~ year + bac08 + bac10 + perse +
    sbprim + sbsecon + sl70plus + gdl + perc14_24log + unemlog +
    vehicmilespclog, data = df.panel, model = "pooling")

model.fe <- plm(totfatrte ~ year + bac08 + bac10 + perse + sbprim +
    sbsecon + sl70plus + gdl + perc14_24log + unemlog + vehicmilespclog,
    data = df.panel, model = "within")

xtable.plm <- xtable:::xtable.lm
```

```
tab1 <- xtable(model.pooled)
print(tab1, size = "\\fontsize{9pt}{10pt}\\selectfont", comment = FALSE)
```

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	-242.3211	8.7102	-27.82	0.0000
year1981	-2.1443	0.8043	-2.67	0.0078
year1982	-6.5027	0.8199	-7.93	0.0000
year1983	-7.4477	0.8331	-8.94	0.0000
year1984	-6.2619	0.8469	-7.39	0.0000
year1985	-7.0284	0.8633	-8.14	0.0000
year1986	-6.4087	0.8977	-7.14	0.0000
year1987	-7.0167	0.9338	-7.51	0.0000
year1988	-7.2149	0.9809	-7.36	0.0000
year1989	-8.8121	1.0182	-8.65	0.0000
year1990	-9.8416	1.0407	-9.46	0.0000
year1991	-12.0290	1.0636	-11.31	0.0000
year1992	-13.8384	1.0853	-12.75	0.0000
year1993	-13.6367	1.0999	-12.40	0.0000
year1994	-13.1485	1.1222	-11.72	0.0000
year1995	-12.5827	1.1505	-10.94	0.0000
year1996	-14.4709	1.1918	-12.14	0.0000
year1997	-14.6038	1.2212	-11.96	0.0000
year1998	-15.1552	1.2390	-12.23	0.0000
year1999	-14.9702	1.2587	-11.89	0.0000
year2000	-15.1222	1.2802	-11.81	0.0000
year2001	-16.0452	1.2966	-12.37	0.0000
year2002	-16.7772	1.3044	-12.86	0.0000
year2003	-17.1431	1.3145	-13.04	0.0000
year2004	-16.6026	1.3427	-12.37	0.0000
bac08	-2.5503	0.5228	-4.88	0.0000
bac10	-1.3102	0.3853	-3.40	0.0007
perse	-0.6634	0.2901	-2.29	0.0224
sbprim	-0.0724	0.4803	-0.15	0.8802
sbsecon	0.0260	0.4188	0.06	0.9506
sl70plus	3.3220	0.4345	7.64	0.0000
gdl	-0.8859	0.5108	-1.73	0.0831
perc14_24log	2.9876	1.8166	1.64	0.1003
unemlog	5.3474	0.4718	11.33	0.0000
vehicmilespclog	28.1388	0.8716	32.29	0.0000

```
tab2 <- xtable(model.fe)
print(tab2, size = "\\fontsize{9pt}{10pt}\\selectfont", comment = FALSE)
```

We see that the pooled regression yields the same result as the OLS model in the previous section. Comparing the coefficients to the fixed effect results, we see changes for the secondary seatbelt law which becomes negative and for the log of the unemployment rate, which becomes negative as well. Given that the only law variable with a positive coefficient is a speed limit of 70 or more, we can say that, based on the fixed effect model and only looking at the coefficients, changes in traffic laws seem to affect traffic fatalities. However, the significance levels changed, e.g. primary seatbelt law is significant at the 0.1% level, while the law variables speed limits of more than 70, *gdl* and secondary seatbelt law are not significant even at the 10% level. Hence, we cannot find evidence that changes in traffic laws **in general** affect traffic fatalities given our defined significance level of at least 10%.

In the pooled regression we have the composite error term with the unobserved time invariant omitted variables. We would need to assume that all our explanatory variables are uncorrelated with this error term and that no serial correlation is present. This assumption appears not likely to hold as omitted variables like sufficient street lighting in all roads of a state might be related

	Estimate	Std. Error	t-value	Pr(>|t|)
year1981	-1.6147	0.4050	-3.99	0.0001
year1982	-3.5508	0.4244	-8.37	0.0000
year1983	-4.2389	0.4376	-9.69	0.0000
year1984	-4.9049	0.4546	-10.79	0.0000
year1985	-5.4773	0.4750	-11.53	0.0000
year1986	-4.6185	0.5082	-9.09	0.0000
year1987	-5.4930	0.5509	-9.97	0.0000
year1988	-6.1640	0.6029	-10.22	0.0000
year1989	-7.5989	0.6435	-11.81	0.0000
year1990	-7.7346	0.6694	-11.55	0.0000
year1991	-8.4420	0.6851	-12.32	0.0000
year1992	-9.4207	0.7096	-13.28	0.0000
year1993	-9.7472	0.7247	-13.45	0.0000
year1994	-10.2074	0.7462	-13.68	0.0000
year1995	-10.0506	0.7722	-13.02	0.0000
year1996	-10.5597	0.8176	-12.92	0.0000
year1997	-10.8280	0.8470	-12.78	0.0000
year1998	-11.6043	0.8659	-13.40	0.0000
year1999	-11.8208	0.8787	-13.45	0.0000
year2000	-12.4113	0.8921	-13.91	0.0000
year2001	-11.8605	0.8998	-13.18	0.0000
year2002	-11.0550	0.9060	-12.20	0.0000
year2003	-11.0933	0.9140	-12.14	0.0000
year2004	-11.6024	0.9376	-12.38	0.0000
bac08	-1.1801	0.3866	-3.05	0.0023
bac10	-0.8965	0.2637	-3.40	0.0007
perse	-1.1568	0.2289	-5.05	0.0000
sbprim	-1.1023	0.3365	-3.28	0.0011
sbsecon	-0.2486	0.2471	-1.01	0.3147
sl70plus	0.2036	0.2652	0.77	0.4429
gdl	-0.4636	0.2865	-1.62	0.1059
perc14_24log	2.6140	1.4133	1.85	0.0646
unemlog	-3.4431	0.3864	-8.91	0.0000
vehicmilespclog	12.5048	1.1430	10.94	0.0000

with some of the variables like unemployment rates (via infrastructure budget). We can test this via the `pooltest()` function.

```
pooltest(model.pooled, model.fe)
```

```
##
##  F statistic
##
## data:  totfatrte ~ year + bac08 + bac10 + perse + sbprim + sbsecon +  ...
## F = 74.333, df1 = 47, df2 = 1118, p-value < 2.2e-16
## alternative hypothesis: unstability
```

We see that the null hypothesis of stability is rejected. In addition, we can test for the presence of serial correlation.

```
pbgtest(model.pooled)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel
##  models
##
## data:  totfatrte ~ year + bac08 + bac10 + perse + sbprim + sbsecon +     sl70plus + gdl + pe
```

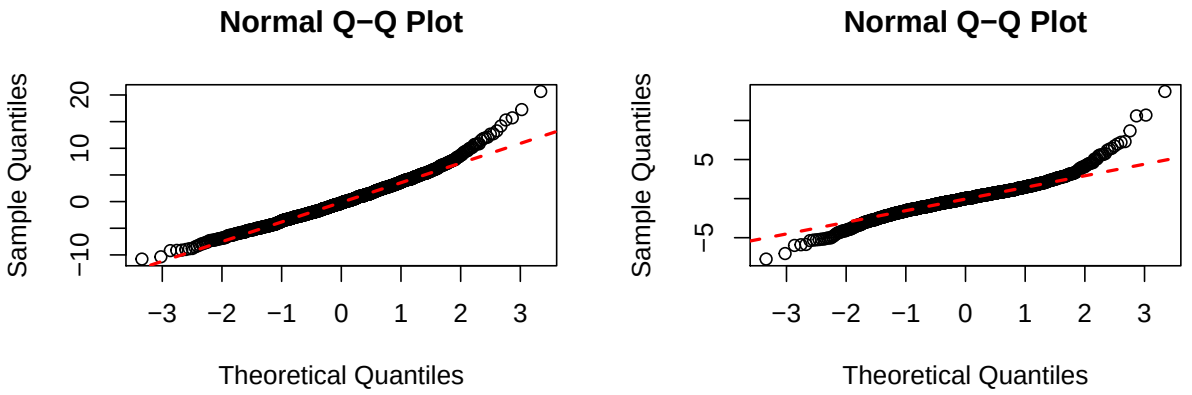
```
## chisq = 775.92, df = 25, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

We see that the null hypothesis of no serial correlation is rejected.

Hence, a pooled model cannot be applied as the asusmptions are violated.

Looking at the residuals for the pooled regression, we see that they deviate from a normal distribution. However, the same holds for the fixed effect model.

```
res.model.pooled <- residuals(model.pooled)
res.model.fe <- residuals(model.fe)
par(mfrow = c(1, 2))
qqnorm(res.model.pooled)
qqline(res.model.pooled, col = 2, lwd = 2, lty = 2)
qqnorm(res.model.fe)
qqline(res.model.fe, col = 2, lwd = 2, lty = 2)
```



In the fixed effect model we just need to assume that no time variant omitted variables are present. Hence, the assumptions under the fixed effect model appear more plausible and consequently we would rather put our trust in the fixed effect model estimates rather than the pooled regression results.

Random or Fixed Effects?

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

Answer:

The disadvantage of a fixed effect model is that the substraction of the mean removes constant or near constant explanatory variables. For example, some laws might always have been present in certain stage. The effect of these laws is removed in the model. Random effects preserve the association with some of the constant explanatory variables in certain states, like seatbelt laws. They require, however, the stronger assumption that the unobserved effect is uncorrelated with all explanatory variables.

We can test whether a random effects model is suitable using the Hausman test. The null hypothesis

is the random effects model being applicable.

```
model.re <- plm(totfatrte ~ year + bac08 + bac10 + perse + sbprim +
    sbsecon + sl70plus + gdl + perc14_24log + unemlog + vehicmilespclog,
    data = df.panel, model = "random")

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

```
##
##  Hausman Test
##
## data:  totfatrte ~ year + bac08 + bac10 + perse + sbprim + sbsecon +  ...
## chisq = 13.332, df = 34, p-value = 0.9994
## alternative hypothesis: one model is inconsistent
```

As the p-value is very high, the null hypothesis of the random effects model being applicable cannot be rejected. Hence, for this dataset a random effects model could be used. Indeed, when looking at the regression results, the coefficients and statistical significance levels of the traffic law variables in the random effect model are similar in magnitude to the fixed effect model.

Fixed Effects Prediction

Exercise: 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.

Answer:

The fixed effects model looks at the difference of the explanatory variables relative to the mean of the variable over time. The predicted effect is the change in the dependent variable *totfatrte*.

As we transformed *vehicmilespc* using the natural log, we need to transform the increase of 1,000 of *vehicmilespc* relative to a baseline by using the natural log as well. As baseline we choose the mean of *vehicmilespc*. We note that the mean is approximately 10,000 - meaning that the increase of 1,000 is roughly a 10% increase over the baseline.

With this in mind, we now predict the effect.

```
increase = 1000
# We need a baseline to compare the increase to, as we use a
# log transformed variable!
base_vehiclemilespc = mean(df_adj$vehicmilespc)
base_vehiclemilespc
```

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

```
change_in_fatality <- exp(model.fe$coefficients["vehicmilespclog"] *
    (log(base_vehiclemilespc + increase) - log(base_vehiclemilespc)))

change_in_fatality
```

```
## vehicmilespclog
##        3.668641
```

Holding all other explanatory variables constant, an increase of 1000 in *vehiclemilespc* relative to the mean level of 9129 is predicted to result in a change of 3.67 in total fataility rate per 100,000. At the overall mean of fatality rate, this would be a 20% increase in fatality rate (so a 10% increase in *vehicmilespc* vs its mean resulted in a 20% increase *totratefte* vs its mean). While this effect is large, it is in line with our observations that the coefficient for *vehicmilespclog* is the largest in the model *and* the variable itself has the largest magnitudes among the included variables. These two facts combined suggest predicted *totfatrte* is relatively sensitive to changes in *vehicmilespclog*.

Consequences of Heteroskedasticity in Idiosyncratic Errors

Exercise: 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?

Answer:

The consequence of serial correlation or heteroskedasticity in the idiosyncratic errors of the model is that ordinary least squares regression and pooled regression would not be reliable, as they assume independent and identically distributed observations. This assumption would be violated with serial correlation (independence) and heteroskedasticity (identically distributed). Instead, fixed effect, random effect or first difference models should be used.

Conclusion

In this project we investigated the question: **"Do changes in traffic laws affect traffic fatalities?"**. We noted that different models, like pooled regression and fixed effects, yield different results. However, the assumptions of ordinary least squares regression or pooled regression are not met as especially serial correlation is present. When looking at fixed and random effects models, we see based on the Hausman test that a random effects model could be used. However, both models provide no evidence that changes in **all** traffic laws do affect traffic fatalities for a defined significance level of at least 10%. Whether a change in a traffic law affects fatality rates as well as the magnitude of the effect depends on **the individual traffic law**.