Lab 3: Panel Models

US Traffic Fatalities: 1980 - 2004

Aruna Bisht, Deepika Maddali

U.S. traffic fatalities: 1980-2004

In this lab, we are asking you to answer the following **causal** question: > **"Do changes in traffic laws affect traffic fatalities?"**

(30 points, total) Build and Describe the Data

1. (5 points) Load the data and produce useful features. Specifically:
   - Produce a new variable, called `speed_limit` that re-encodes the data that is in `sl55`, `sl65`, `sl70`, `sl75`, and `slnone`;
   - Produce a new variable, called `year_of_observation` that re-encodes the data that is in `d80`, `d81`, `...` , `d04`.
   - Produce a new variable for each of the other variables that are one-hot encoded (i.e. `bac*` variable series).
   - Rename these variables to sensible names that are legible to a reader of your analysis. For example, the dependent variable as provided is called, `totfatrte`. Pick something more sensible, like, `total_fatalities_rate`. There are few enough of these variables to change, that you should change them for all the variables in the data. (You will thank yourself later.)

Answer. In order to answer the research question, we created useful features through encoding. We encoded the four speed limit columns into one, named 'speed_limit'. Similarly, all year columns (number 25) were combined into a single column called "year of observation". Likewise, the blood alcohol limit columns with values 0.08 and 0.1 were combined into one column. This encoding approach reduces the number of columns compared to creating separate dummy variables for each individual column. The renaming of the columns was done for the sake of clarity.

2. (5 points) Provide a description of the basic structure of the dataset. What is this data? How, where, and when is it collected? Is the data generated through a survey or some other method? Is the data that is presented a sample from the population, or is it a *census* that represents the entire population? Minimally, this should include and How is the our dependent variable of interest `total_fatalities_rate` defined?

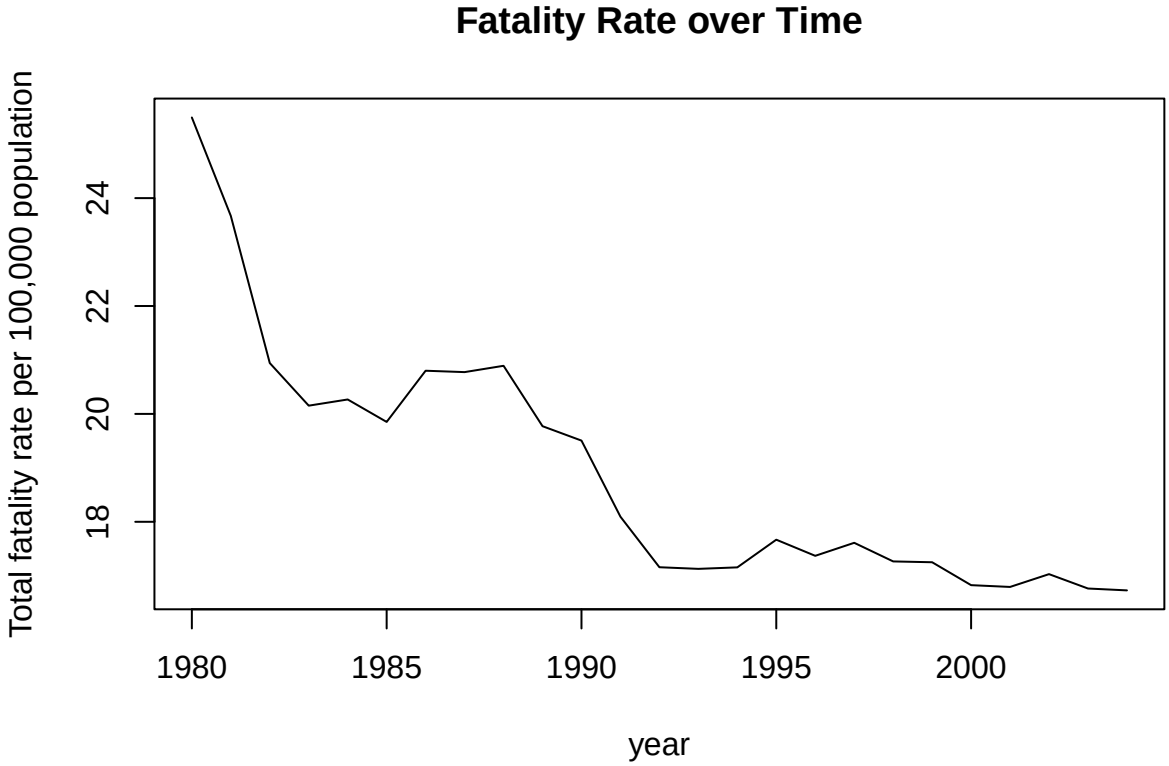
Answer. This dataset follows a panel data structure, tracking information for the 48 US states over the years 1980 to 2004. With this structure, we can analyze both within-state and between-state variables. The dataset covers various driving-related aspects, encompassing changes in drunk driving laws, seat belt usage, speed limits, blood alcohol limits, and economic and demographic factors. It comprises 56 columns, representing observations for all 48 states from 1980 to 2004.

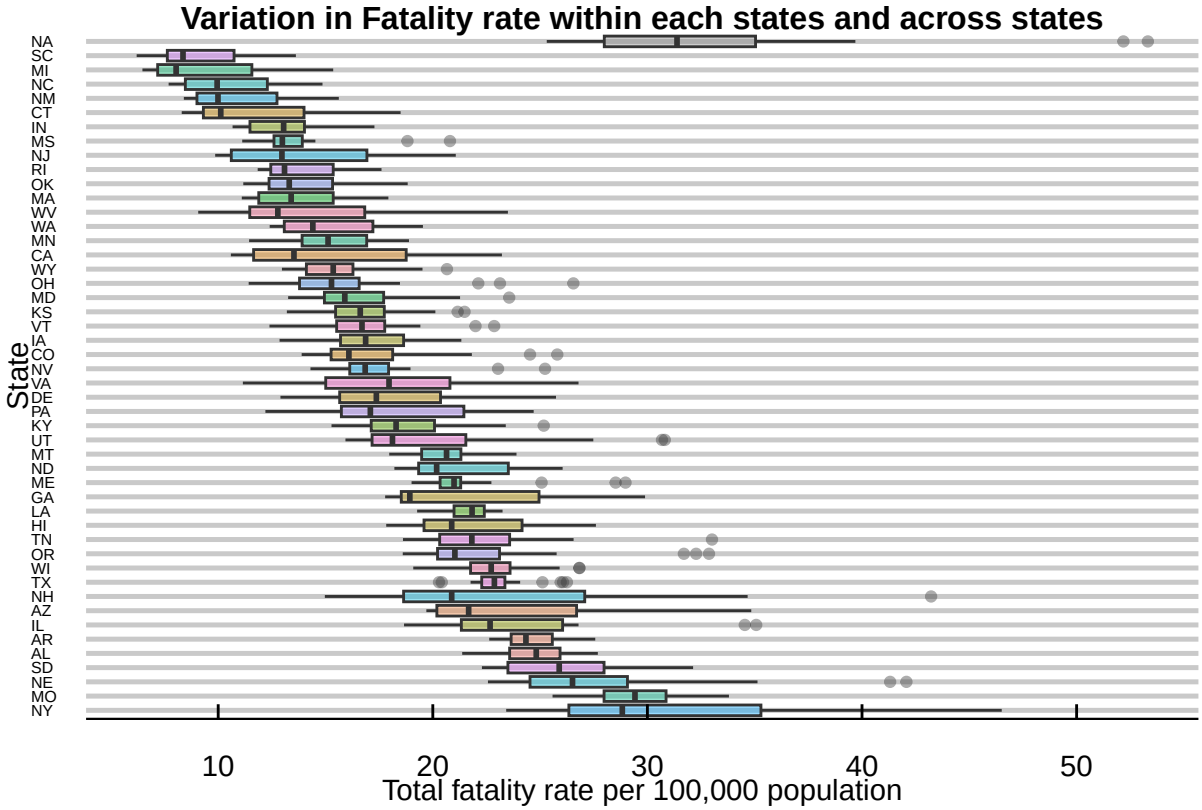
The data is collected through a survey, providing valuable insights into traffic-related trends and patterns. Although it does not represent the entire population of the United States, it serves as a sample from the population due to its focus on the 48 states. Therefore, any findings derived from this sample can be generalized to the larger population of these contiguous states.

The dependent variable of interest in this dataset is the 'total fatalities rate,' defined as the total number of fatalities occurring per 100,000 population. In panel data, it is common to use rates as dependent or response variables, as this enables us to account for variations in exposure or population size over time and across different entities in the dataset. This approach ensures that comparisons and interpretations are more meaningful and accurate, taking into consideration the specific characteristics of each state and the entire time period under study.

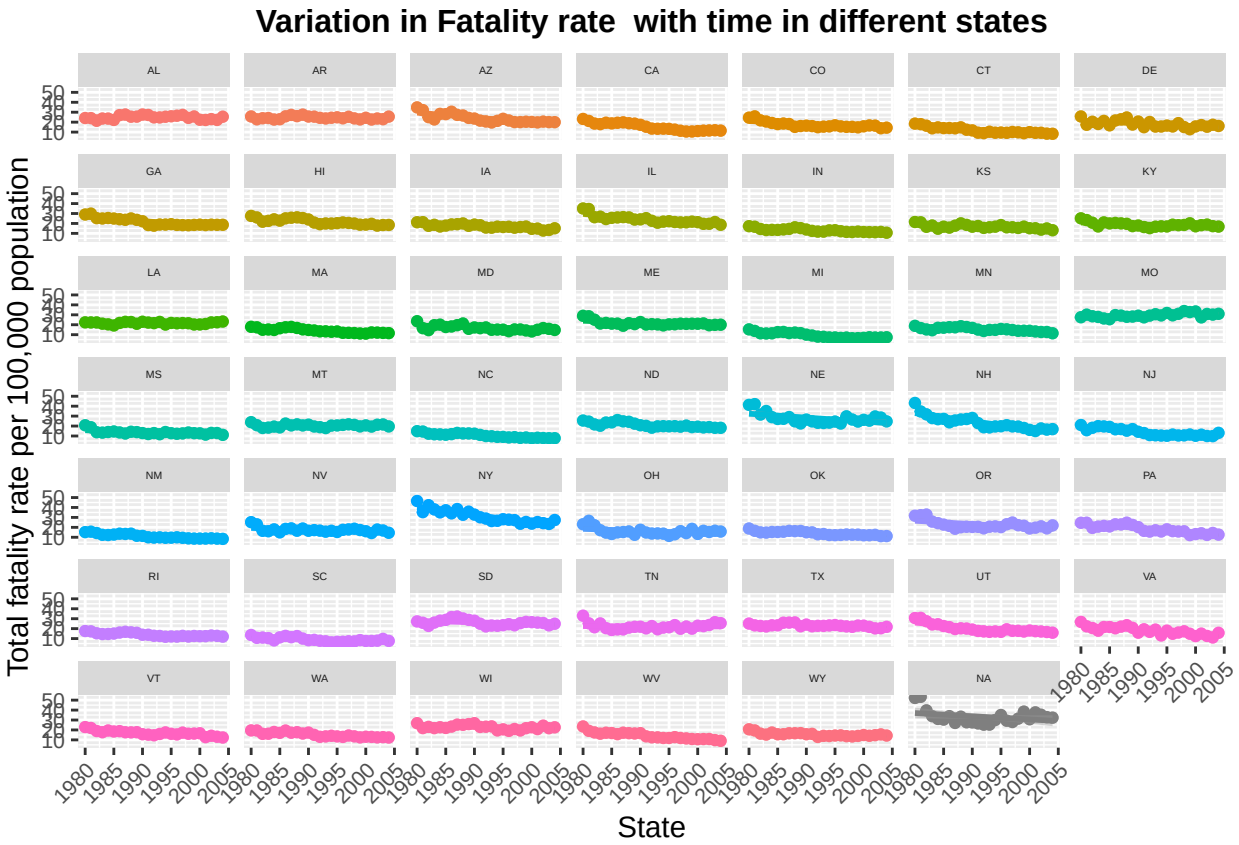
3 (20 points) Conduct a very thorough EDA, which should include both graphical and tabular techniques, on the dataset, including both the dependent variable `total_fatalities_rate` and the potential explanatory variables. Minimally, this should include: How is the our dependent variable of interest `total_fatalities_rate` defined? and What is the average of `total_fatalities_rate` in each of the years in the time period covered in this dataset?

Answer. The dependent variable of interest in this dataset is the 'total fatalities rate,' defined as the total number of fatalities occurring per 100,000 population. By taking this dependent variable as rate, it allows us to accommodate fluctuations in exposure or population size across diverse entities within the dataset over time. The plot displays the average total_fatalities_rate for each year, demonstrating a decreasing trend over time. The max fatality rate is in year 1980 with value 25.49458. The min fatality rate is in year 2004 with value 16.72896

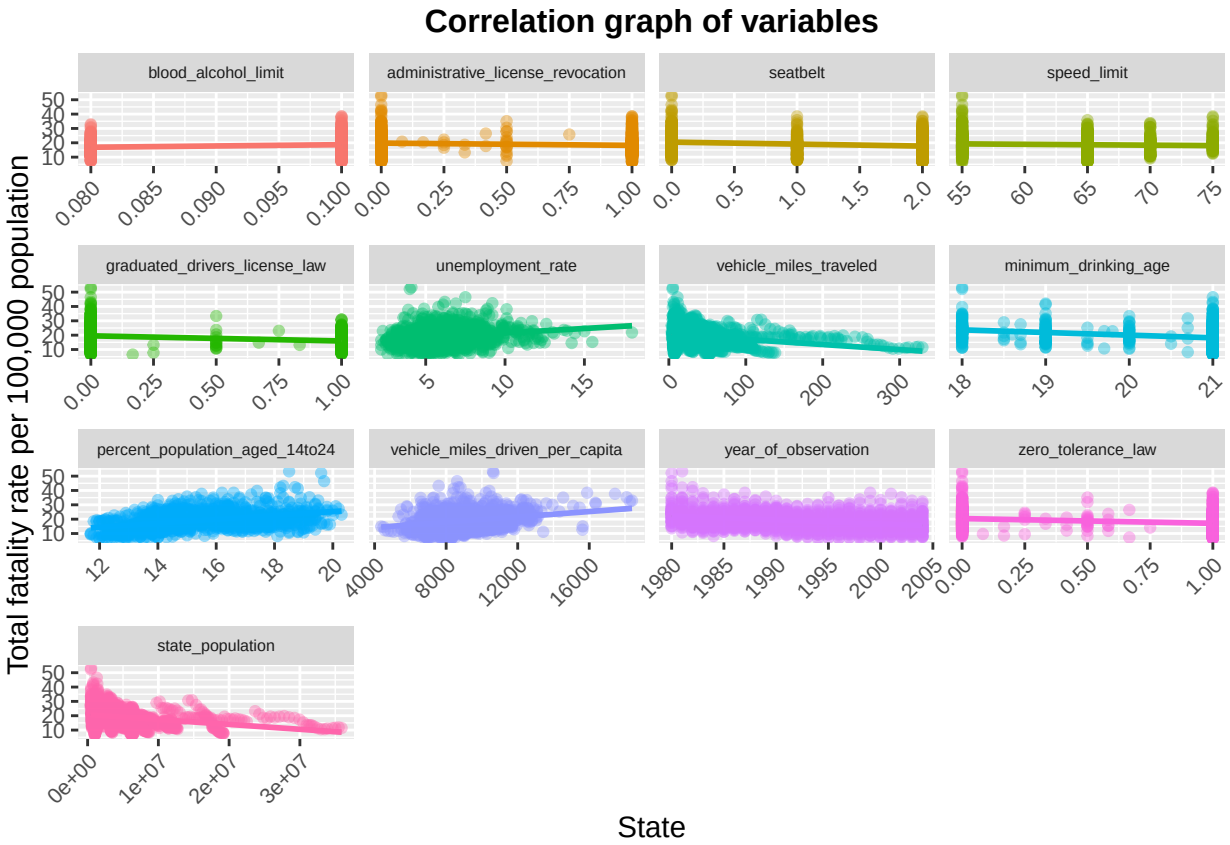




>**Answer.** We see strong differences in fatality rate across states and there is good variation of fatality rate within the states as well, suggesting that fixed effects are important for controlling for unobserved differences.



>**Answer.** We see that states appear to have diverged in fatality rate over time, with some between more fatality rate and others shifting to less fatality rate.

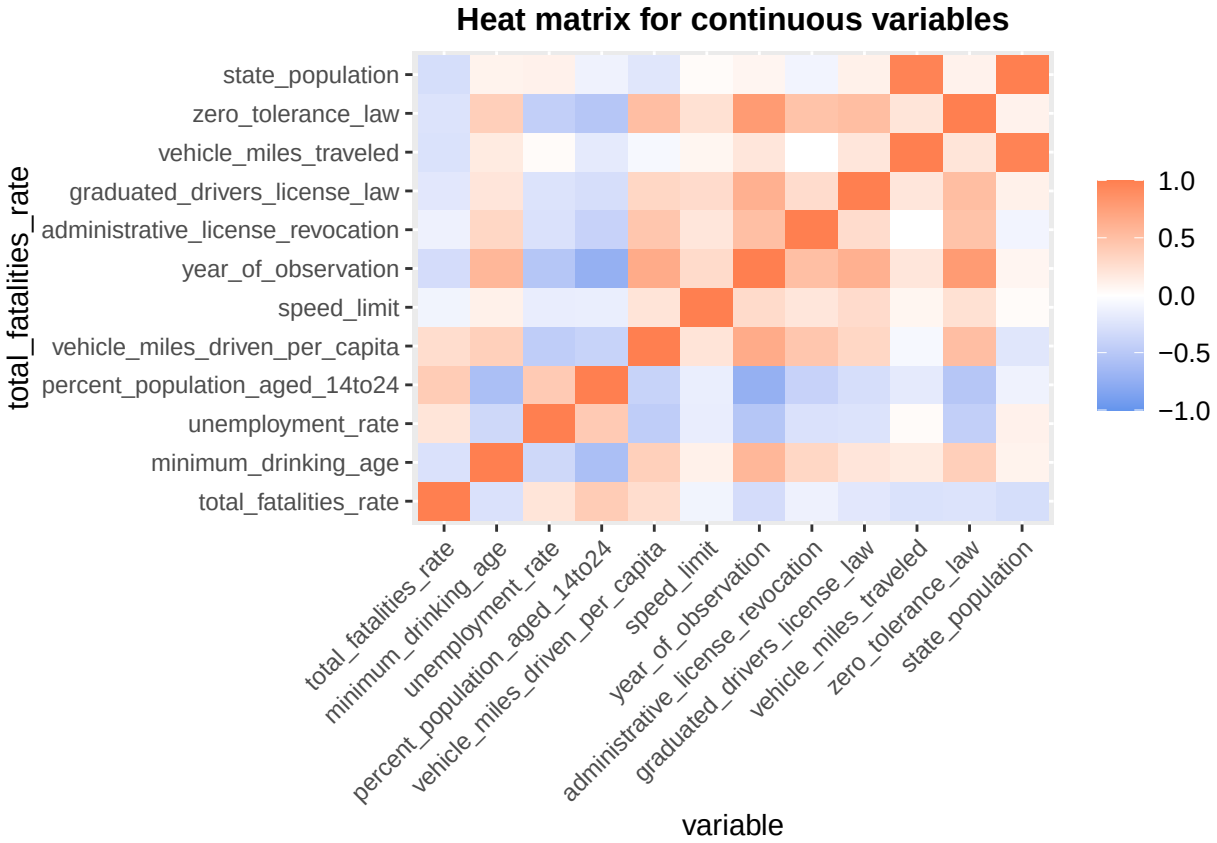


Answer. Regarding explanatory variables, some variables, such as adiminstrative_license_revocation, seatbelt, speed limit, graduated driving license law, the percentage of the population between 14 and 24 years old, the unemployment rate, vehicle miles driven per capita, and blood alcohol limit appear to strongly affect the fatality rate and show an overall upward trend with the fatality rate. Hence included in the model as explanatory variables. Variables like minimum drinking age, zero tolerance law don't have an obviously strong overall impact and exhibit a downward trend with the fatality rate. Therefore, they have been excluded from the model. Variable 'state population' is a demograhic variable which is indirectly included if we take state as fixed effects. Hence variable 'state population' is excluded from the model.Similarly, the variable 'vehicle_miles_traveled' was excluded from the model because we already have a similar variable, vehicle miles driven per capita, in the explanatory variables. Including both would result in redundant information and potentially cause multicollinearity issues.

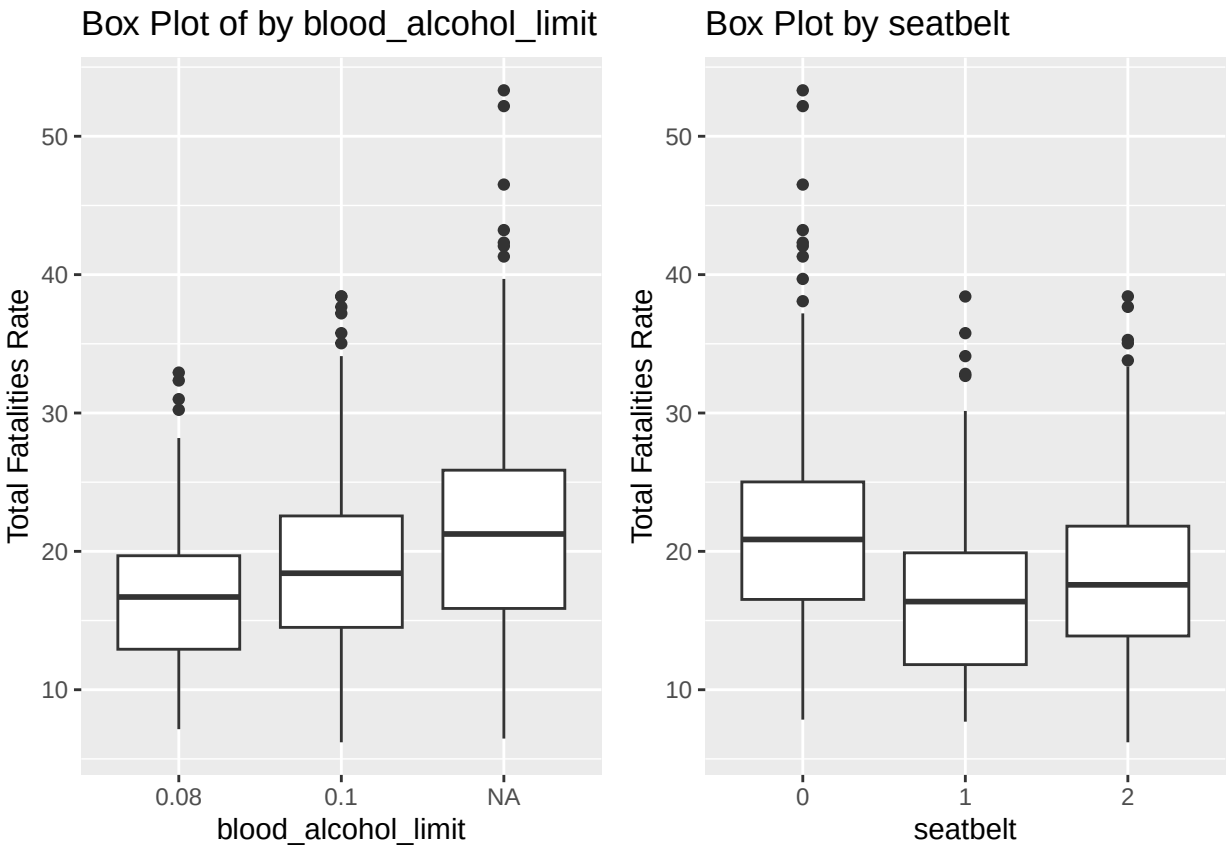
Our response variable is total_fatalities_rate, so we cannot include similar fatality variables like totalfat, nightfat, wikndfat, totfatpvm, nightfatpvm, wkndfatpvm, nghtfatrte, wkndfatrte, along with other explanatory variables in the model equation. This is because it might lead to a strong relationship with other independent variables, leading to multicollinearity, inflated standard errors, and unreliable estimates. Instead, we use the response variable total_fatality_rate per 100,000 population, as it represents the overall fatality rate, instead of focusing solely on nighttime or weekend fatalities. Additionally, we prefer using total_fatality_rate per 100,000 population over total_fatality_rate per million miles because we are interested in how traffic laws implemented on the population affect fatality rates. The former variable is better suited for this purpose as it directly relates to the impact of traffic laws on public safety. Also we chose total_fatality_rate per 100,000 population over other variable name 'total fatality' because when comparing fatality rates between different states with varying population sizes, using rates per 100,000 population standardizes the comparison and provides a fair assessment of the effectiveness of traffic laws.

This plot in the appendix categorizes variables into discrete and continuous values. We observed that 'blood alcohol limit' and 'seatbelt' have discrete values. So they were treated as categorical variables. On the other hand, the rest of the variables exhibit a continuous distribution of frequency, making them suitable for

classification as continuous variables in the dataset. Some variable 'administrative license revocation' and 'graduated driving license law' do not exceed 1, suggesting that they can be interpreted as percentages of the population.



Answer. Certain varaibles like vehicle_miles_driven_per_capita and administrative_license_revocation seem highly related to one another. Similary vehicle_miles_driven_per_capita and unemployment rate is also correlated with each other. While others like the percentage of population between 14&24 and speed limit are not related. Similarly, according to the correlation matrix of categorical variables shown in the appendix, certain variables like blood alcohol limit and seat belt appear to be highly correlated with each other.




```
## NULL
```

This shows relation between categorical variable blood_alcohol_limit and response variable fatality rate. There are a higher number of 'NA' values compared to the other values, which were taken as zero. The median of 0.1 is higher than 0.08, indicating a higher fatality rate. By looking at the size of the boxes, there is more variability in the fatality rate at blood alcohol limit 0.1 compared to blood alcohol limit 0.08. Similary relation between categorical variable seat belt and response variable fatality rate shows there is a higher number of '0' values compared to the other values for seat belt. The median of 2 is slightly higher than 1, indicating a slightly higher fatality rate. By looking at the size of the boxes, there is more or less the same variability in the fatality rate between seat belt values 1 and 2.

(15 points) Preliminary Model

Estimate a linear regression model of *totfatrte* on a set of dummy variables for the years 1981 through 2004 and interpret what you observe. In this section, you should address the following tasks:

- Why is fitting a linear model a sensible starting place?.

Answer. Fitting a linear model is a sensible place to start because it allows us to explore the correlations between independent variables. Linear models serve as a baseline to establish relationships with the initial variable and provide a foundation for building more complex models on top of it. Additionally, linear models are simple and easy to interpret

- What does this model explain, and what do you find in this model?.

Answer. This model captures the time effects within our panel dataset. All the time estimates in this model are negative, implying that the fatality rate from 1981 onwards was lower compared to the baseline fatality rate in 1980. For instance, the fatality rate in 1981 was 1.824 units lower than the rate in 1980.

By employing year as dummy variables, this model adeptly manages time effects and effectively captures substantial variations in the response variable.

- Did driving become safer over this period? Please provide a detailed explanation.

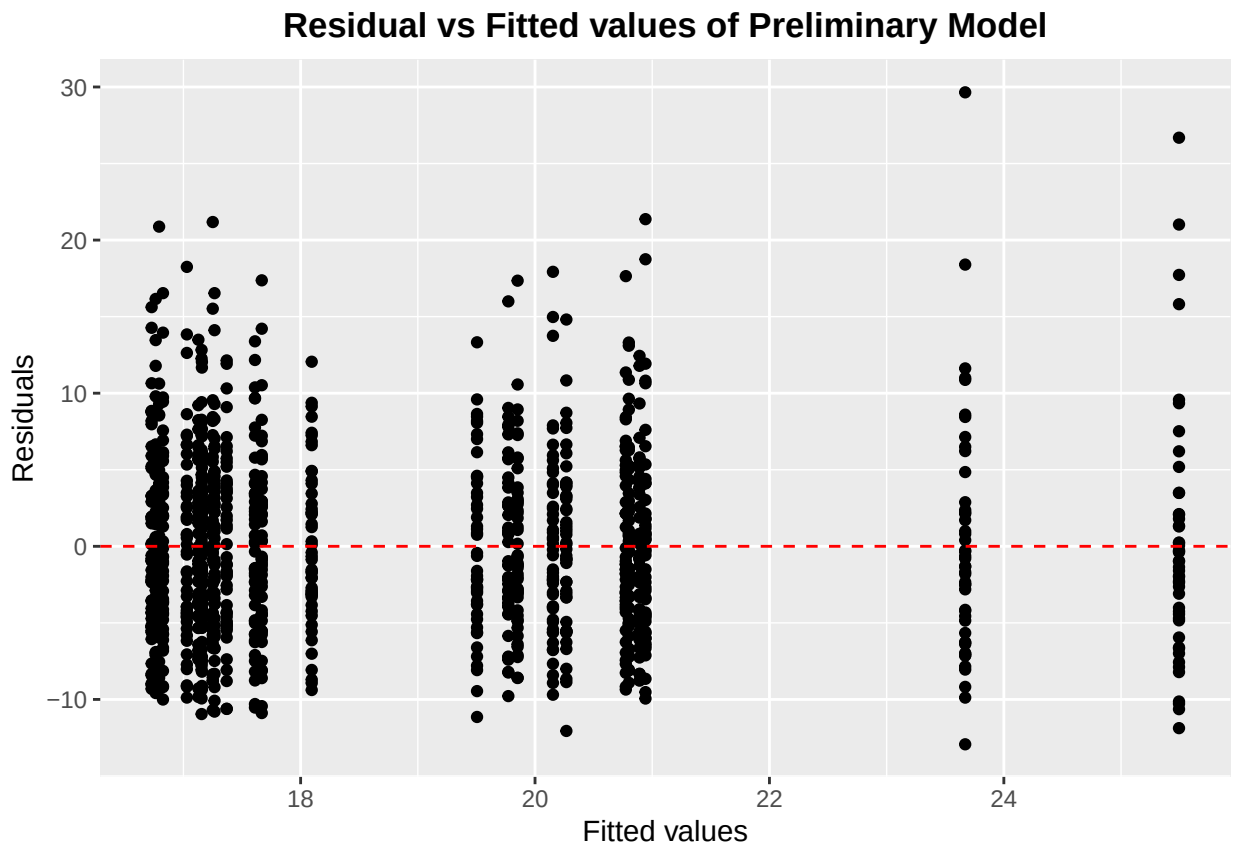
Answer. Yes, driving became safer over time, as evident from the summary table of the model. Year 1980 is dropped off and taken as the baseline year in model. The magnitude of each coefficient represents the size of the estimated change in the total_fatalities_rate associated with that year compared to the reference year 1980 and the overall values of time coefficient in the model is decreasing with time. We also observed same in the EDA previosly that the fatality rate is decreasing over the period of time from 1998 to 2000. All years are statistically signifincat and impact fatality rate except year 1981. There is a slight increase in the fatality rate from year 1985 to 1987 but overrall it has a downward trend.

- What, if any, are the limitation of this model. In answering this, please consider **at least**:
  - Are the parameter estimates reliable, unbiased estimates of the truth? Or, are they biased due to the way that the data is structured?
  - Are the uncertainty estimate reliable, unbiased estimates of sampling based variability? Or, are they biased due to the way that the data is structured?

Answer. Our dataset is a panel dataset with observations spanning both time and states. This model captures and fixes the variation in time effects. However, a limitation of this preliminary model is that it doesn't account for time-invariant state effects, thus failing to consider the unobserved heterogeneity across states. This unaccounted heterogeneity could result in omitted bias and introduce bias to the current estimates. Furthermore, the presence of this omitted variable bias can lead to underestimated standard errors, potentially leading to incorrect conclusions regarding the statistical significance of our coefficients. As evident from the residual plot, a pattern is noticeable, suggesting that unexplained variation still exists in the response data, further contributing to biased estimates. Yes,uncertainty estimates can be reliable due to sample variability and randomness.

```
model_fe_residuals <- residuals(lm_preliminary_model)

ggplot(data.frame(fitted_values = fitted(lm_preliminary_model) , residuals = model_fe_residuals) , aes(
```



(15 points) Expanded Model

Expand the **Preliminary Model** by adding variables related to the following concepts:

- Blood alcohol levels
- Per se laws
- Primary seat belt laws (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.)
- Secondary seat belt laws
- Speed limits faster than 70
- Graduated drivers licenses
- Percent of the population between 14 and 24 years old
- Unemployment rate
- Vehicle miles driven per capita.

If it is appropriate, include transformations of these variables. Please carefully 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.

Answer. Yes, a few transformations were needed in the data. The variable 'year of observation' was converted into a factor to treat it as a categorical variable. This allowed us to assess the impact of each year on the fatality rate. The variable 'blood alcohol limit' contained some missing values (NA) in the dataset. It was important to replace these missing values with 0 before treating it as a categorical variable. This enabled us to measure the impact on the fatality rate when the blood alcohol limit was increased from 0 to 0.08 and from 0 to 0.1.Similarly, the variable 'Seat belt' was also converted into a categorical variable because it is a descrete variable. This allowed us to understand the impact on fatality if a person transitioned from not wearing a seatbelt to wearing a primary seatbelt and then from not wearing a seatbelt to wearing a secondary

seatbelt. The rest of the variables were treated as numerical.The response variable 'total_fatalities_rate' is skewed, as confirmed by histogam and the Shapiro test in appendix. To address this skewness and avoid biased estimates, a log transformation was applied to achieve a more symmetric distribution.

- How are the blood alcohol variables defined? Interpret the coefficients that you estimate for this concept.

Answer. The variable 'Blood alcohol limit' was treated as categorical. It appears that both levels of blood alcohol limit have a significant impact on the fatality rate. If there is a unit increase in the blood alcohol limit from 0 to 0.08 then fatality lower by 0.93 units in comparison to baseline fatality when BAC limit is 0. The p value is 0.0124 so this BAC 0.08 limit is statistically significant and have an impact on fatality rate. However BAC 0.1 limit is not statistically significant and has a P value of 0.1521. This may be because evidence for its effect is weaker. Or this could be because the model is not complex enough to accurately capture the relationship between total fatality rate and blood alcohol limit 0.1.

- Do *per se laws* have a negative effect on the fatality rate?

Answer. Variable 'per se' has a negative effect on the fatality rate but this variable did not come significant . And this is probably because again the model is not complex to capture true relationship of this variable with response variable.

- Does having a primary seat belt law?

Answer. The variable 'primary seat belt' exhibits a negative impact on the fatality rate; however, this variable did not achieve statistical significance. This lack of significance is likely attributed to the model's simplicity, which might not adequately capture the true relationship between this variable and the response variable.

(15 points) State-Level Fixed Effects

Re-estimate the **Expanded Model** using fixed effects at the state level.

- What do you estimate for coefficients on the blood alcohol variables? How do the coefficients on the blood alcohol variables change, if at all?

Answer. The change in coefficients for both blood alcohol limits, 0.08 and 0.1, is that they both have now become significant. This significance is attributed to the improved capturing of the relationship within this fixed effect model. As the blood alcohol limit increases from 0 to 0.08, the fatality rate decreases by 0.9702 units in comparison to the baseline fatality rate when the BAC limit is 0. The P value for BAC 0.08 is 0.02, indicating a significant impact on the fatality rate. Similarly, the P value for BAC 0.1 is even lower, at 0.005, suggesting higher significance. This makes sense, as we would expect the fatality rate to decrease further as the blood alcohol limit increases. Specifically, when the blood alcohol limit changes from 0 to 0.1, the fatality rate decreases by 0.9726 units compared to the baseline fatality rate when the BAC is 0.

- What do you estimate for coefficients on per se laws? How do the coefficients on per se laws change, if at all?

Answer. The coefficient of Administrative license revocation is also significant, with a P-value of 8.667e-09. It has a negative estimator, suggesting that its presence decreases the fatality rate. A unit increase in administrative license revocation is associated with a decrease in the fatality rate by 0.8135 units in comparison to the baseline fatality rate when there is no administrative per se law.

- What do you estimate for coefficients on primary seat-belt laws? How do the coefficients on primary seatbelt laws change, if at all?

Answer. The primary seat belt is also significant with a P-value of 0.002027. It has a significant impact in comparison to the secondary seat belt. A unit increase in the presence of the primary seat belt would decrease the fatality rate by 0.8478 units compared to the baseline fatality rate when there is no seat belt.

Which set of estimates do you think is more reliable? Why do you think this? **Answer.**AAmong the Expanded model and the Fixed model, the estimates obtained from the Fixed effect model are more reliable.

This is because the Fixed effect model accounts for state-level fixed effects, capturing unobserved heterogeneity across states and addressing omitted variable bias.

- What assumptions are needed in each of these models?
  Answer. The assumptions needed to apply the model using State-Level Fixed effects are as follows:

Assumption 1. Fixed effect model satisfy this equation

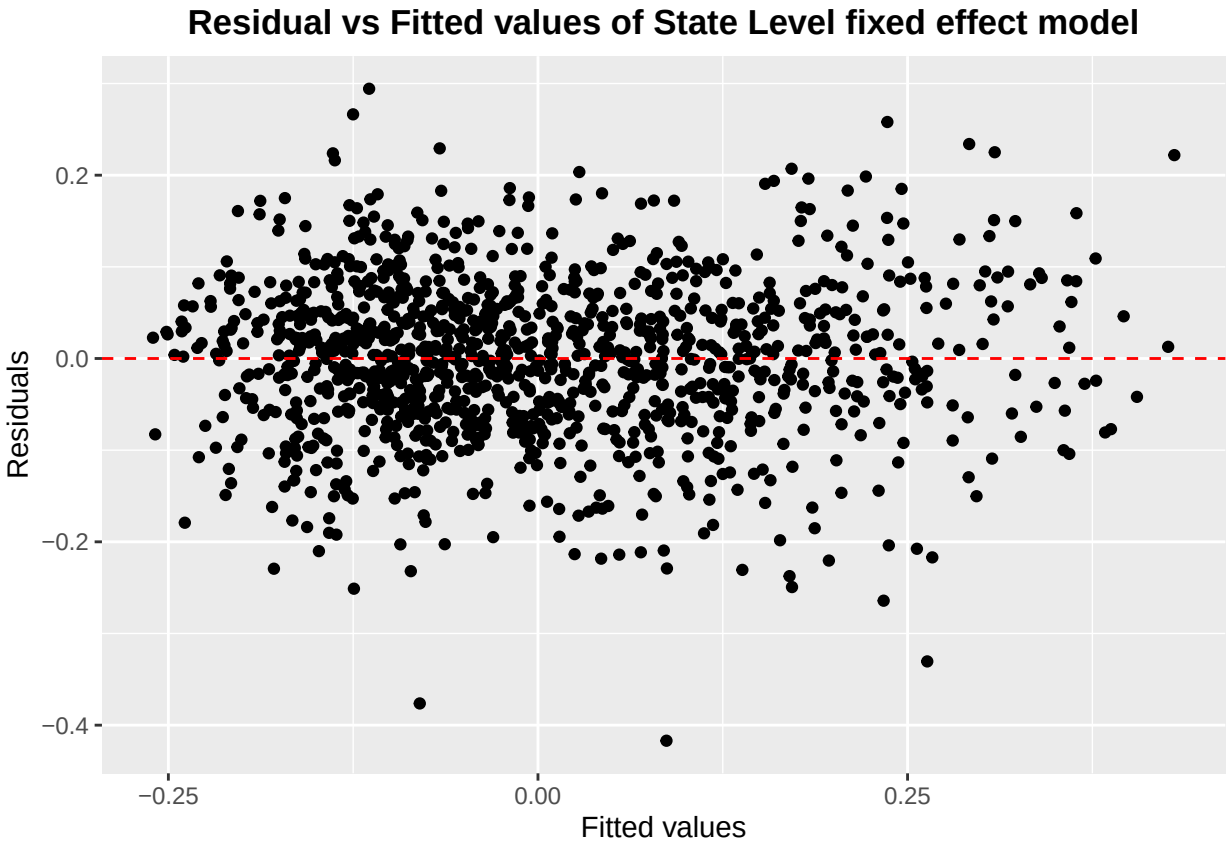
$$y_{it} = \beta_0 + \beta_1 x_{1it} + ... + \beta_p x_{pit} + a_i + \epsilon_{it}$$

In a Fixed Effect model, the term ai represents the time-invariant individual-specific effect, which can capture unobserved heterogeneity or factors that do not change over time. In our dataset, we have state-specific characteristics, such as geography, that remain constant over time and can influence the outcome variable, i.e., fatality rate. Therefore, the Fixed Effect model is appropriate.

Assumption 2: We have a random sample from the cross-section. There are 57 observations for 47 states over a period of 25 years. While this particular assumption is not fully satisfied, our model has accounted for the major sources of variation that are most relevant for control purposes.

Assumption 3: This assumption states that each explanatory variable changes over time for at least some individuals in the sample. In our dataset, we have 57 observations that vary over a period of 25 years for 47 states

Assumption 4: E(epsilon_it | Xi, ai) = 0 . As observed from the residual plot, the residuals do not exhibit any pattern. This indicates that the assumption of no correlation between errors and explanatory variables, given the time-invariant effect, is satisfied.



Assumption 5: Var(epsilon_it | Xi , ai) = Var(epsilon_it) = sigmaˆ2 , for all t = 1,2,3,. . . ,T However, based on the Breusch Pagan test for homoscedasticity, we obtained a p-value less than 0.05. This indicates that errors have serial correlation, and therefore, the homoscedasticity assumption is not satisfied

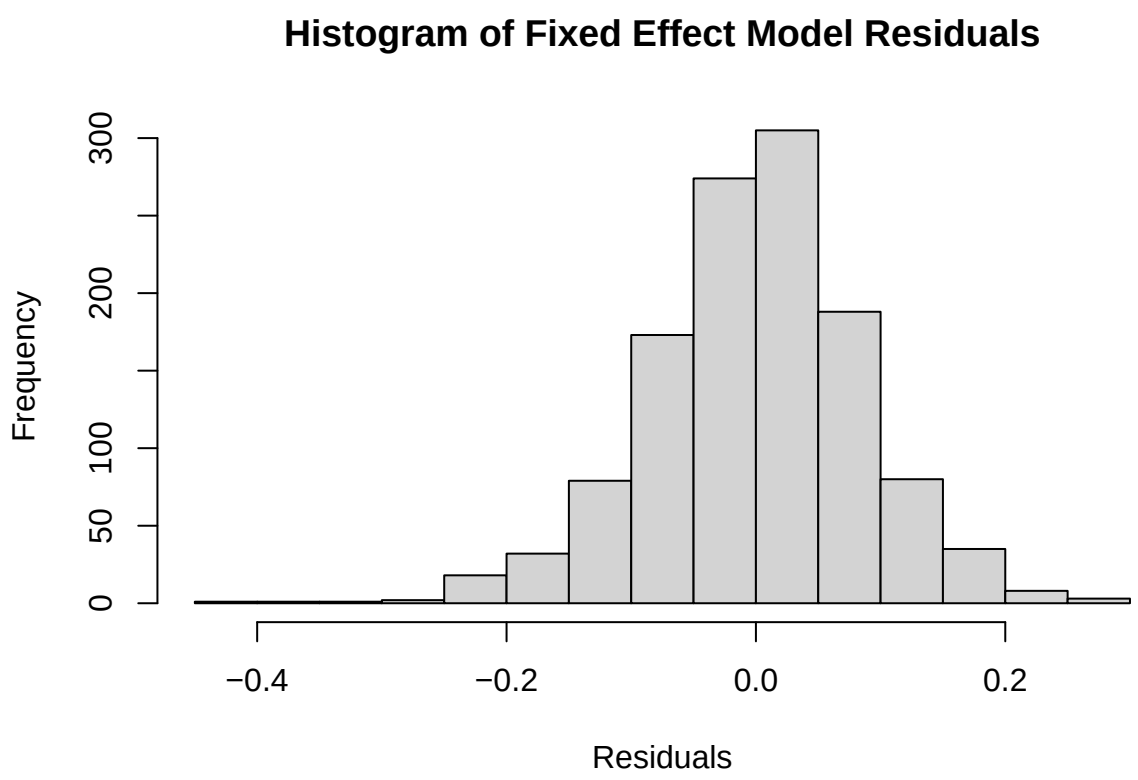
```
##
##  Breusch-Pagan LM test for cross-sectional dependence in panels
##
## data:  log(total_fatalities_rate) ~ year_of_observation + blood_alcohol_limit +     administrative_li
## chisq = 2799.5, df = 1128, p-value < 2.2e-16
## alternative hypothesis: cross-sectional dependence
```

Assumption 6: Cov(epsilon_it, epsilon_is | Xi, ai) = 0. The assumption of uncorrelated errors is typically met when idiosyncratic errors are uncorrelated, given all explanatory variables and ai. However, in our case, there is evidence of serial correlation because Durbin Watson test is significant with p value 2.2e-16, indicating a deviation from the assumption of uncorrelated errors. To rectify this issue and accommodate heteroskedasticity, a different type of error treatment is required.In addressing the concern of serial correlation within residuals, R offers two primary options. Arellano standard errors and Newey-West errors both provide the solution that addresses both group heteroskedasticity and serial correlation by considering temporal correlation within groups which we would calculate in the last section.

```
##
##  Durbin-Watson test for serial correlation in panel models
##
## data:  log(total_fatalities_rate) ~ year_of_observation + blood_alcohol_limit +  ...
## DW = 1.1828, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel models
##
## data:  log(total_fatalities_rate) ~ year_of_observation + blood_alcohol_limit +  ...
## chisq = 218.2, df = 2, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

Assumption 7: Normal(0, u^2). As seen from the histogram the plot of residuals is normal.Hence this assumption is satisfied.



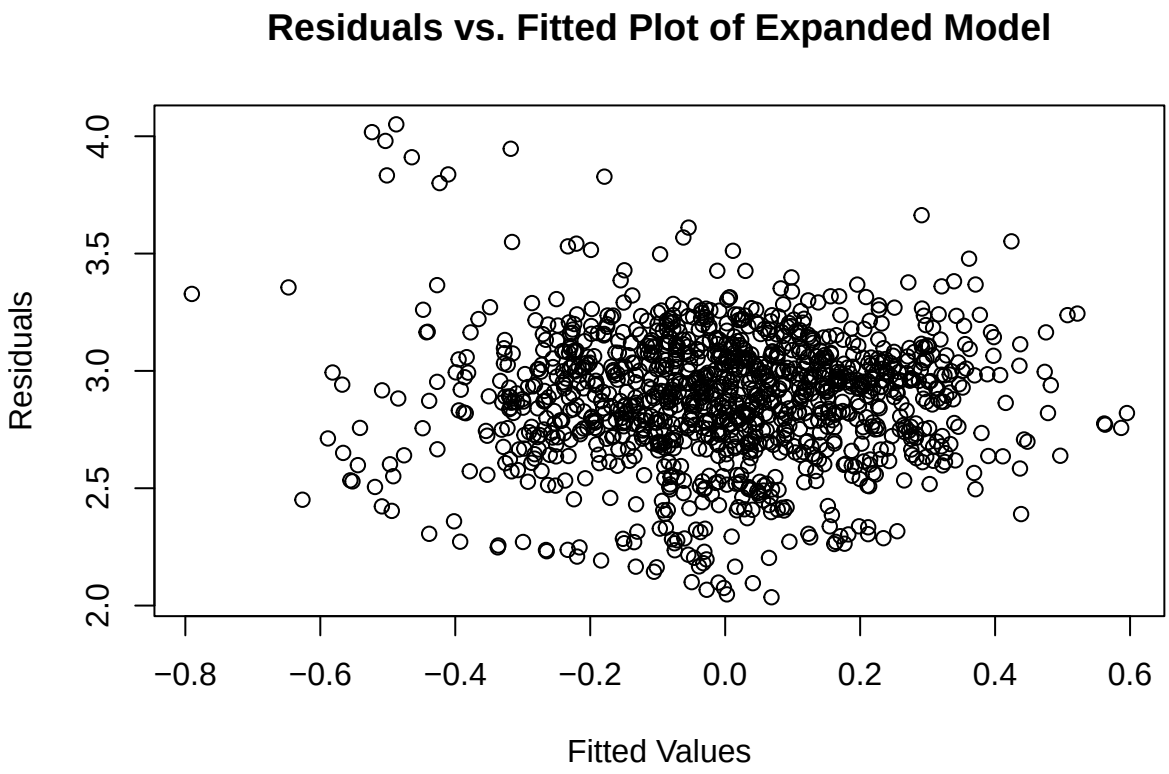
The Multiple Linear Expanaded model Assumptions assumption are: Answer. Assumption 1: Linear in parameters

$$y_{it} = \beta_0 + \beta_1 x_{1it} + ... + \beta_p x_{pit} + a_i + \epsilon_{it}$$

Assumption 2: Random Sampling - The goal is to have a random sample of n observations, following the population model as per the first assumption. However, in the case of the multiple linear regression equation, this particular assumption is not satisfied. Nevertheless, the model attempts to account for the variation by including time fixed effect as year intercept and other explanatory variables.

Assumption 3: No Perfect Collinearity - To assess this assumption, we examined the heat matrix and found a few correlations between explanatory variables. However, they were not significantly high enough to indicate perfect collinearity

Assumption 4: Zero Conditional Mean - When plotted with the explanatory variable, there is no clear pattern of residuals. However, there seems to be a slight clustering in between. E(epsilon_it | Xi) = 0



Assumption 5: Homoskedasticity - the error term (mu) has the same variance given any value of the explanatory variable, as seen from this plot. There is a slight cluster of errors in the center. Var(epsilon_it | Xi) = sigma^2

- Are these assumptions reasonable in the current context?

Answer. In the present context, the assumptions of the fixed effect model appear to be more reasonable when compared to the assumptions of multiple linear regression. These fixed effect assumptions hold true in the provided dataset, as different states exhibit distinct individual effects that remain consistent over time. This accounts for the state-level heterogeneities that could potentially influence the fatality rate.

However, it's important to note that these fixed effect assumptions might not always be valid, contingent on the specific situation. For instance, the assumption that state-level effects remain constant over time may not hold if there are significant changes in state-level policies or other time-varying factors that could impact the fatality rate. Similarly, the presence of omitted variables that have a relationship with the independent variables can lead to bias. Therefore, while the fixed effect model assumptions are appropriate in this scenario, their validity should be assessed considering the potential influences of various factors and variables.

```
##
##  Durbin-Watson test
##
## data:  lm_expanded_model
## DW = 0.32792, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

(10 points) Consider a Random Effects Model

Instead of estimating a fixed effects model, should you have estimated a random effects model?

- Please state the assumptions of a random effects model, and evaluate whether these assumptions are met in the data.

Answer. The random effect model shares many similarities with the fixed effect model, but there is one key distinction.

Assumption 1: In the random effect model, the critical assumption is that the expected value of ai (the individual-specific effect) given all explanatory variables (Xi) is constant: E(ai | Xi) = beta. This means there is no correlation between the unobserved effect and the explanatory variables, which distinguishes it from the fixed effect model.

To assess this assumption, we can employ a statistical Hausman test. The test's purpose is to examine whether the residuals in the random effects model are uncorrelated with other predictors in the model, indicating the absence of omitted variable bias. The null hypothesis states that the random effects model is appropriate, while the alternative hypothesis suggests correlation between residuals and other predictors, favoring the fixed effects model to remove omitted variable bias.

Upon conducting the test, we got the p-value less than 0.05, so we reject the null hypothesis of the random effects model and instead apply the fixed effects model. This fixed model allows for the correlation between individual-specific effects and explanatory variables, violating the random effect model assumption.

Assumption 2: Cov(epsilon_it, epsilon_is | Xi, ai) = 0. For all t != s, the idiosyncratic errors are uncorrelated (conditional on all explanatory variables and ai). Since there is a serial correlation so would have to take a different kind of error to take into account heteroscadscity .

```
##
##  Hausman Test
##
## data:  log(total_fatalities_rate) ~ year_of_observation + blood_alcohol_limit +  ...
## chisq = 912.91, df = 33, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

```
##
##  Durbin-Watson test for serial correlation in panel models
##
## data:  total_fatalities_rate ~ year_of_observation + blood_alcohol_limit +  ...
## DW = 0.93025, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel models
##
## data:  total_fatalities_rate ~ year_of_observation + blood_alcohol_limit +  ...
## chisq = 374.09, df = 2, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

Assumption 3: Var(epsilon_it | Xi , ai) = Var(epsilon_it) = sigmaˆ2 , for all t = 1,2,3,. . . ,T As per the homoscadescity test name Beauchan test, we got the p value less then 0.05

```
##
##  Breusch-Pagan LM test for cross-sectional dependence in panels
##
## data:  log(total_fatalities_rate) ~ year_of_observation + blood_alcohol_limit +     administrative_li
## chisq = 2799.5, df = 1128, p-value < 2.2e-16
## alternative hypothesis: cross-sectional dependence
```

- If the assumptions are, in fact, met in the data, then estimate a random effects model and interpret the coefficients of this model. Comment on how, if at all, the estimates from this model have changed compared to the fixed effects model.

Answer. Since Haufman test failed so Random effects model assumptions are not met. So even if we consider the estimates of the Random effect model then it would be bias because as per the Haufman test the correlation exist between error and independent variable.

** If the assumptions are **not** met, then do not estimate the data. But, also comment on what the consequences would be if you were to *inappropriately* estimate a random effects model. Would your coefficient estimates be biased or not? Would your standard error estimates be biased or not? Or, would there be some other problem that might arise?**

Answer. Since the Fixed effect assumptions are not met so we would not estimate the data. The random effects model assumes that the random effects are uncorrelated with the independent variables. However, if this assumption is violated and if we still inappropriately estimate the random effect model, the coefficient estimates can be biased. The random effects model assumes that the random effects are uncorrelated with the error term but if this assumption is violated and if we still estimate the random effect model, the standard error estimates can be biased, leading to incorrect assessments of the statistical significance of coefficients.Violation of assumptions can lead to invalid statistical inference. Confidence intervals and hypothesis tests may not be accurate, potentially leading to incorrect conclusions about the significance of variables.

(10 points) Model Forecasts

The COVID-19 pandemic dramatically changed patterns of driving. Find data (and include this data in your analysis, here) that includes some measure of vehicle miles driven in the US. Your data should at least cover the period from January 2018 to as current as possible. With this data, produce the following statements:

- Comparing monthly miles driven in 2018 to the same months during the pandemic:
  - What month demonstrated the largest decrease in driving? How much, in percentage terms, lower was this driving? **Answer.**The monthe which demonstrated the largest decrease driving is Jan month of 2020. In percentage terms, driving was lower by -6.5% in comparision to 2018
  - What month demonstrated the largest increase in driving? How much, in percentage terms, higher was this driving? **Answer.**The monthe which demonstrated the largest increase driving is April month of 2020. In percentage terms, driving was higher by 39.07% in comparision to 2018.

Now, use these changes in driving to make forecasts from your models. - Suppose that the number of miles driven per capita, increased by as much as the COVID boom. Using the FE estimates, what would the consequences be on the number of traffic fatalities? Please interpret the estimate. **Answer.**Utilizing fixed-effect estimates, the projected traffic fatalities in April 2020, during the COVID-19 pandemic, would be elevated by 107,525 in terms of the fatality rate. Essentially, we extracted the coefficient of vehicle miles per capita from the fixed-effect model and then multiplied it by the total increase in vehicle miles during the COVID-19 period. The mileage derived from the COVID dates is on a per capita basis.

- Suppose that the number of miles driven per capita, decreased by as much as the COVID bust. Using the FE estimates, what would the consequences be on the number of traffic fatalities? Please interpret the estimate. **Answer.**Utilizing fixed-effect estimates, the projected traffic fatalities in January 2020, during the COVID-19 pandemic, would be reduced by -16,111 in terms of the fatality rate. Similarly, we extracted the coefficient of vehicle miles per capita from the fixed-effect model and then multiplied it by the total decrease in vehicle miles during the COVID-19 period. The mileage derived from the COVID dates is on a per capita basis.

‘

```
max_increase_row <- joined_data %>%
  filter(max_increase == max(max_increase))
max_increase_row
```

```
## # A tsibble: 1 x 13 [1D]
##   DATE.x     vehicle_miles_2018 month DATE.y     vehicle_miles_2020 DATE
##   <date>                  <dbl> <dbl> <date>                  <dbl> <date>
```

```
## 1 2018-04-01             275127     4 2020-04-01             167617 2021-04-01
## # i 7 more variables: vehicle_miles_2021 <dbl>, diff_2018_2020 <dbl>,
## #   diff_2018_2021 <dbl>, max_increase <dbl>, max_decrease <dbl>,
## #   max_increase_percentage <dbl>, max_decrease_percentage <dbl>
```

```
## [1] 107510
```

```
## # A tsibble: 1 x 13 [1D]
##   DATE.x     vehicle_miles_2018 month DATE.y     vehicle_miles_2020 DATE
##   <date>                  <dbl> <dbl> <date>                  <dbl> <date>
## 1 2018-01-01             244736     1 2020-01-01             260847 2021-01-01
## # i 7 more variables: vehicle_miles_2021 <dbl>, diff_2018_2020 <dbl>,
## #   diff_2018_2021 <dbl>, max_increase <dbl>, max_decrease <dbl>,
## #   max_increase_percentage <dbl>, max_decrease_percentage <dbl>
```

```
## Fatality rate in the month 4 has the maximum number of fatalities with 6.575603 fatalities.
```

```
## Fatality rate in the month 1 has the maximum number of fatalities with -0.9853925 fatalities.
```

(5 points) Evaluate Error

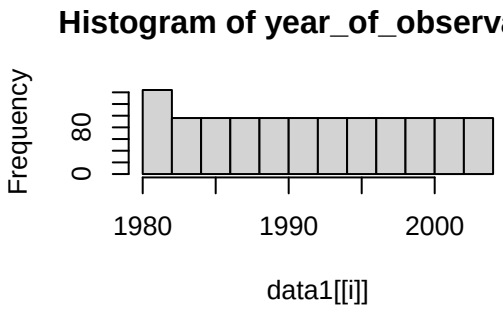
If there were serial correlation or heteroskedasticity in the idiosyncratic errors of the model, what would be the consequences on the estimators and their standard errors? Is there any serial correlation or heteroskedasticity?
Answer. If there is a serial correlation in the idiosyncratic errors of the model then the estimators would still be unbias & consistent but it can make them less efficient. However, Standard errors would be underestimated, leading to incorrect inference about the signifincance of estimated coefficients. The p value may be too small then expected leading to unlikely significance of variable.

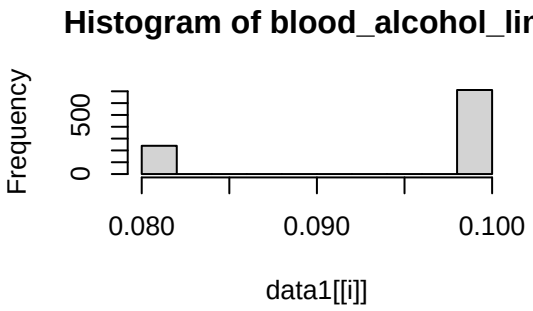
If there is a heteroskedasticity in the idiosyncratic errors of the model then the estimators would be unbias & consistent but it can make them less efficient, leading to less precise estimates. However, the standard errors would be bias leading to unreliable inferencing. If the standard errors are not account for these situations then standard error can be either too large or too small leading to unreliable hypothesis testing.

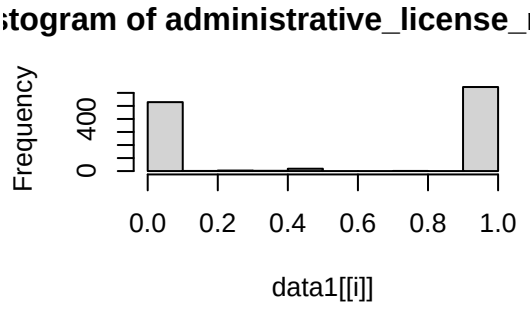
Yes there is a serial correlation and heteroskedasticity in the fixed effect model .In order to address this concern, R offers two primary options Arellano standard errors and Newey-West. They provide a solution that addresses both group heteroskedasticity and serial correlation by considering temporal correlation within groups. To calculate errors, we employ the 'vcovHC' function, which offers a range of types: white1 (robust standard errors), white2 (cluster-robust standard errors), Arellano (incorporating temporal correlation within groups), and Newey-West (using 'vcovNW'). Hence Arellano and Newey-West choice offers enhanced robustness due to its ability to account for correlation across states/groups. Consequently, these standard errors are slightly larger than regular OLS errors, a trade-off for increased efficiency and bias reduction resulting from addressing these complexities.

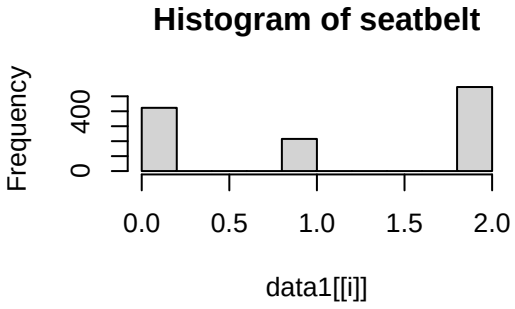
```
##             Type         SE
## 1    Regular OLS 0.01818164
## 2         Robust 0.02449075
## 3 Cluster Robust 0.01755873
## 4     Newey West 0.02122974
## 5      Arrellano 0.01703066
```

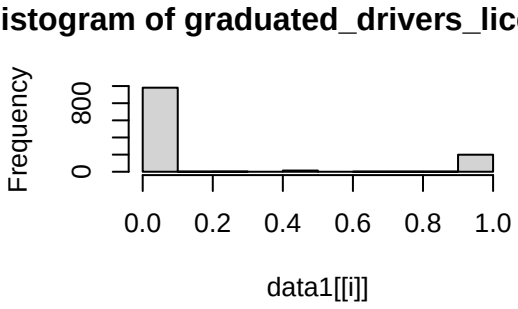
Appendix

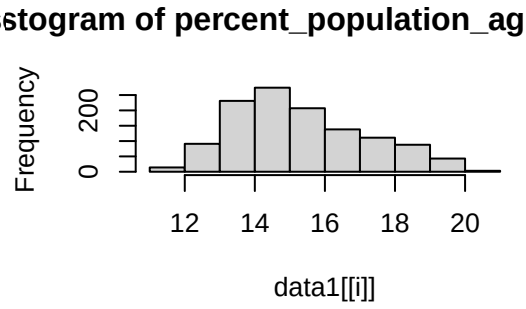

Histogram of year_of_observation
Frequency
0
80
1980
1990
2000
data1[[i]]

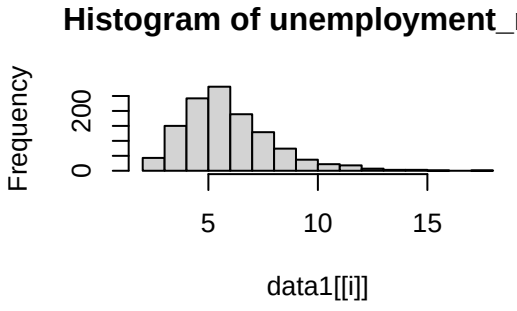

Histogram of blood_alcohol_limit
Frequency
0
500
0.080
0.090
0.100
data1[[i]]

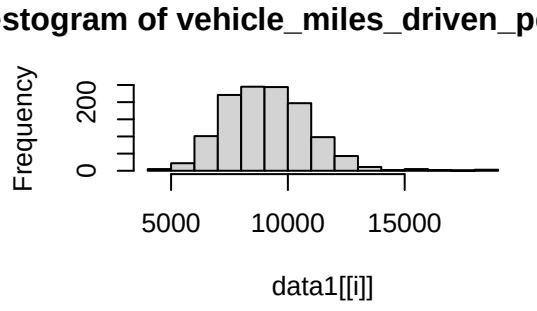

stogram of administrative_license_revo
Frequency
0
400
0.0
0.2
0.4
0.6
0.8
1.0
data1[[i]]

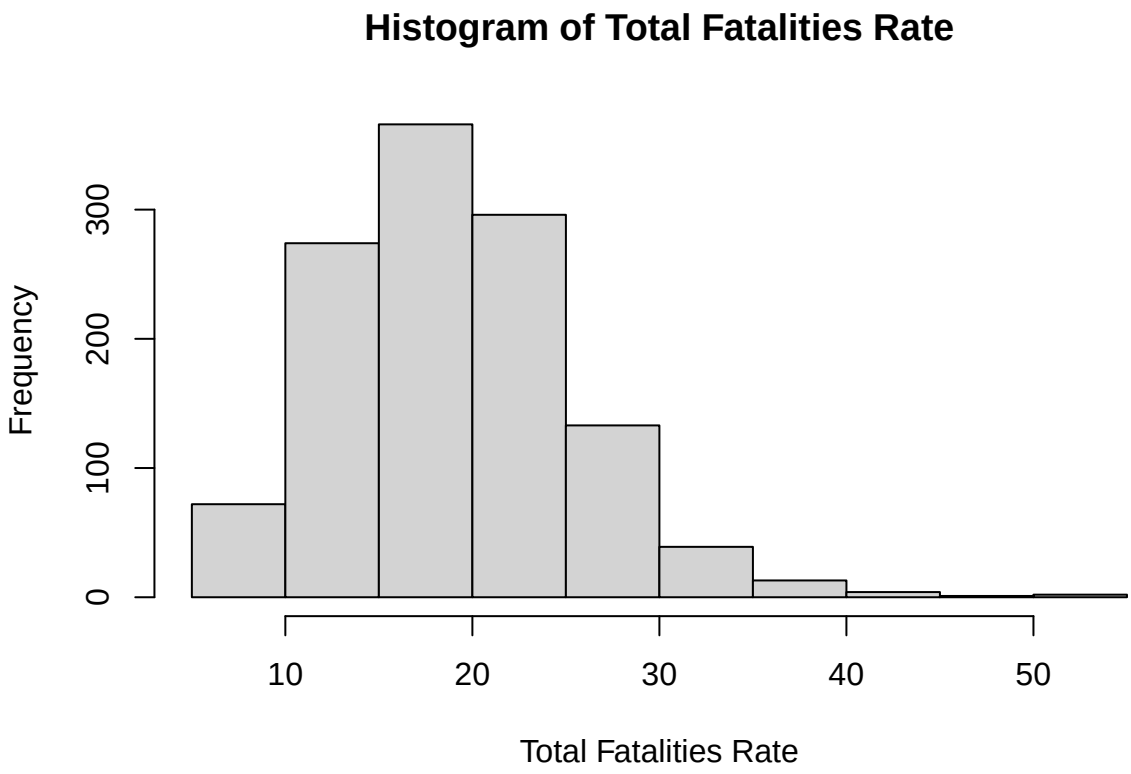

Histogram of seatbelt
Frequency
0
400
0.0
0.5
1.0
1.5
2.0
data1[[i]]


istogram of graduated_drivers_licens
Frequency
0
800
0.0
0.2
0.4
0.6
0.8
1.0
data1[[i]]


stogram of percent_population_aged_
Frequency
0
200
12
14
16
18
20
data1[[i]]

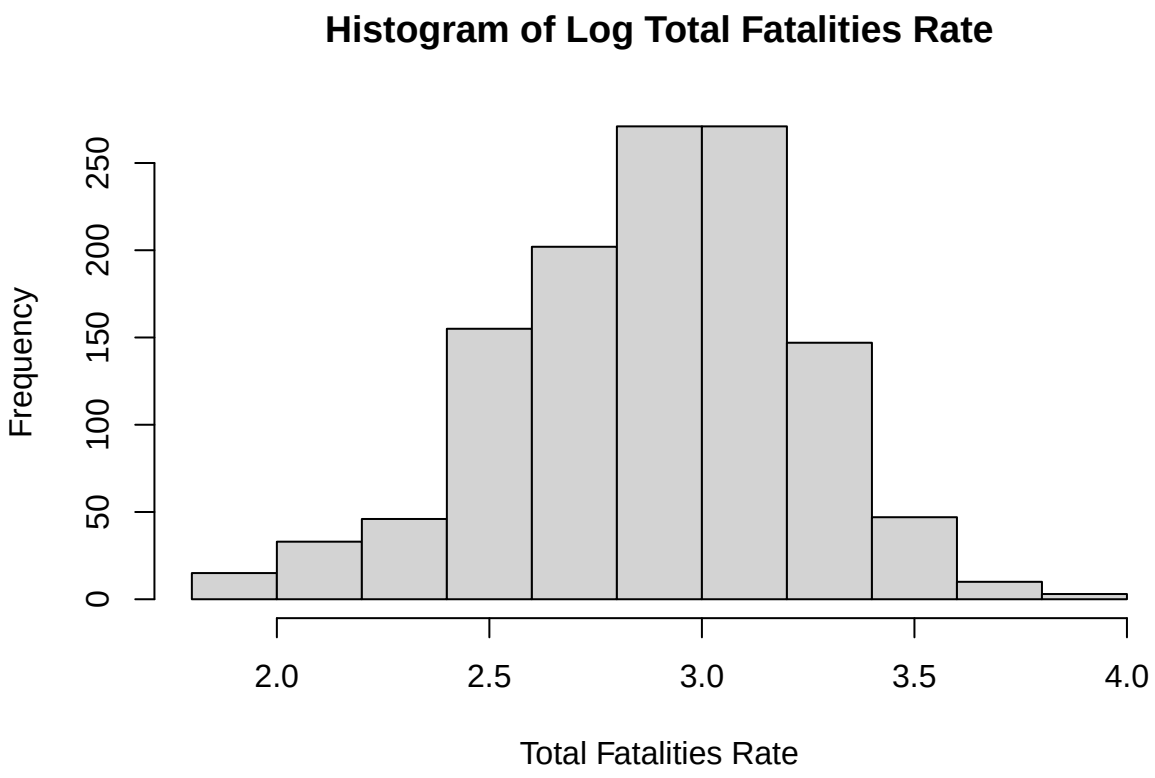

Histogram of unemployment_rate
Frequency
0
200
5
10
15
data1[[i]]


stogram of vehicle_miles_driven_per_
Frequency
0
200
5000
10000
15000
data1[[i]]




```
##
##  Shapiro-Wilk normality test
##
## data:  data1$total_fatalities_rate
## W = 0.96832, p-value = 1.572e-15
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(data1$total_fatalities_rate)
## W = 0.99093, p-value = 9.448e-07
```




```
# Preliminary Model of *totfatrte* on a set of dummy variables for the years 1981 through 2004
lm_preliminary_model <- lm(total_fatalities_rate ~  d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89
summary(lm_preliminary_model)
```

```
##
## Call:
## lm(formula = total_fatalities_rate ~ d81 + d82 + d83 + d84 +
##     d85 + d86 + d87 + d88 + d89 + d90 + d91 + d92 + d93 + d94 +
##     d95 + d96 + d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04,
##     data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9302  -4.3468  -0.7305   3.7488  29.6498
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  25.4946     0.8671  29.401  < 2e-16 ***
## d81          -1.8244     1.2263  -1.488 0.137094
## d82          -4.5521     1.2263  -3.712 0.000215 ***
## d83          -5.3417     1.2263  -4.356 1.44e-05 ***
## d84          -5.2271     1.2263  -4.263 2.18e-05 ***
## d85          -5.6431     1.2263  -4.602 4.64e-06 ***
## d86          -4.6942     1.2263  -3.828 0.000136 ***
## d87          -4.7198     1.2263  -3.849 0.000125 ***
## d88          -4.6029     1.2263  -3.754 0.000183 ***
```

```
## d89          -5.7223     1.2263  -4.666 3.42e-06 ***
## d90          -5.9894     1.2263  -4.884 1.18e-06 ***
## d91          -7.3998     1.2263  -6.034 2.14e-09 ***
## d92          -8.3367     1.2263  -6.798 1.68e-11 ***
## d93          -8.3669     1.2263  -6.823 1.43e-11 ***
## d94          -8.3394     1.2263  -6.800 1.66e-11 ***
## d95          -7.8260     1.2263  -6.382 2.51e-10 ***
## d96          -8.1252     1.2263  -6.626 5.25e-11 ***
## d97          -7.8840     1.2263  -6.429 1.86e-10 ***
## d98          -8.2292     1.2263  -6.711 3.01e-11 ***
## d99          -8.2442     1.2263  -6.723 2.77e-11 ***
## d00          -8.6690     1.2263  -7.069 2.67e-12 ***
## d01          -8.7019     1.2263  -7.096 2.21e-12 ***
## d02          -8.4650     1.2263  -6.903 8.32e-12 ***
## d03          -8.7310     1.2263  -7.120 1.88e-12 ***
## d04          -8.7656     1.2263  -7.148 1.54e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.008 on 1175 degrees of freedom
## Multiple R-squared:  0.1276, Adjusted R-squared:  0.1098
## F-statistic: 7.164 on 24 and 1175 DF,  p-value: < 2.2e-16
```

```
# Expanded Model
lm_expanded_model <- lm(log(total_fatalities_rate) ~   year_of_observation + blood_alcohol_limit +   adı

summary(lm_expanded_model)
```

```
##
## Call:
## lm(formula = log(total_fatalities_rate) ~ year_of_observation +
##     blood_alcohol_limit + administrative_license_revocation +
##     seatbelt + sl70plus + graduated_drivers_license_law + percent_population_aged_14to24 +
##     unemployment_rate + vehicle_miles_driven_per_capita, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79036 -0.13173  0.00297  0.14421  0.59540
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.529e+00  1.276e-01  11.976  < 2e-16 ***
## year_of_observation1981        -9.393e-02  4.270e-02  -2.200   0.0280 *
## year_of_observation1982        -3.019e-01  4.402e-02  -6.859 1.12e-11 ***
## year_of_observation1983        -3.519e-01  4.456e-02  -7.899 6.47e-15 ***
## year_of_observation1984        -2.766e-01  4.489e-02  -6.161 9.92e-10 ***
## year_of_observation1985        -3.058e-01  4.588e-02  -6.665 4.07e-11 ***
## year_of_observation1986        -2.797e-01  4.780e-02  -5.853 6.28e-09 ***
## year_of_observation1987        -3.088e-01  4.967e-02  -6.216 7.10e-10 ***
## year_of_observation1988        -3.193e-01  5.206e-02  -6.133 1.18e-09 ***
## year_of_observation1989        -3.967e-01  5.412e-02  -7.330 4.29e-13 ***
## year_of_observation1990        -4.488e-01  5.538e-02  -8.104 1.33e-15 ***
## year_of_observation1991        -5.608e-01  5.658e-02  -9.913  < 2e-16 ***
## year_of_observation1992        -6.658e-01  5.771e-02 -11.536  < 2e-16 ***
## year_of_observation1993        -6.604e-01  5.834e-02 -11.320  < 2e-16 ***
```

```
## year_of_observation1994           -6.520e-01  5.945e-02 -10.968  < 2e-16 ***
## year_of_observation1995           -6.384e-01  6.083e-02 -10.494  < 2e-16 ***
## year_of_observation1996           -7.615e-01  6.287e-02 -12.113  < 2e-16 ***
## year_of_observation1997           -7.948e-01  6.417e-02 -12.385  < 2e-16 ***
## year_of_observation1998           -8.475e-01  6.507e-02 -13.025  < 2e-16 ***
## year_of_observation1999           -8.612e-01  6.595e-02 -13.058  < 2e-16 ***
## year_of_observation2000           -8.834e-01  6.702e-02 -13.182  < 2e-16 ***
## year_of_observation2001           -9.330e-01  6.802e-02 -13.717  < 2e-16 ***
## year_of_observation2002           -9.659e-01  6.883e-02 -14.034  < 2e-16 ***
## year_of_observation2003           -9.906e-01  6.892e-02 -14.374  < 2e-16 ***
## year_of_observation2004           -9.802e-01  7.068e-02 -13.867  < 2e-16 ***
## blood_alcohol_limit0.08           -5.983e-02  2.390e-02  -2.503   0.0124 *
## blood_alcohol_limit0.1            -2.549e-02  1.779e-02  -1.433   0.1521
## administrative_license_revocation -1.644e-02  1.532e-02  -1.074   0.2832
## seatbelt1                         -3.007e-03  2.529e-02  -0.119   0.9054
## seatbelt2                          2.081e-02  2.215e-02   0.940   0.3476
## sl70plus                           2.331e-01  2.297e-02  10.147  < 2e-16 ***
## graduated_drivers_license_law     -8.107e-04  2.716e-02  -0.030   0.9762
## percent_population_aged_14to24     1.542e-02  6.329e-03   2.437   0.0150 *
## unemployment_rate                  3.829e-02  4.019e-03   9.529  < 2e-16 ***
## vehicle_miles_driven_per_capita    1.577e-04  4.901e-06  32.183  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2087 on 1165 degrees of freedom
## Multiple R-squared:  0.6436, Adjusted R-squared:  0.6331
## F-statistic: 61.86 on 34 and 1165 DF,  p-value: < 2.2e-16
```

```
#  State-Level Fixed Effects
within.model <- plm(log(total_fatalities_rate) ~   year_of_observation + blood_alcohol_limit +   adminis
summary(within.model)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log(total_fatalities_rate) ~ year_of_observation +
##     blood_alcohol_limit + administrative_license_revocation +
##     seatbelt + sl70plus + graduated_drivers_license_law + percent_population_aged_14to24 +
##     unemployment_rate + vehicle_miles_driven_per_capita, data = data2,
##     model = "within", index = c("state", "year_of_observation"))
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -0.4169845 -0.0520969  0.0039513  0.0530077  0.2942811
##
## Coefficients:
##                                    Estimate  Std. Error  t-value  Pr(>|t|)
## year_of_observation1981           -5.9312e-02  1.8182e-02  -3.2622  0.001139
## year_of_observation1982           -1.1776e-01  1.9468e-02  -6.0491 1.984e-09
## year_of_observation1983           -1.4450e-01  1.9941e-02  -7.2460 7.983e-13
## year_of_observation1984           -1.7933e-01  2.0191e-02  -8.8819 < 2.2e-16
## year_of_observation1985           -1.9882e-01  2.1126e-02  -9.4112 < 2.2e-16
## year_of_observation1986           -1.5132e-01  2.2586e-02  -6.6997 3.302e-11
```

```
## year_of_observation1987          -1.8682e-01  2.4267e-02  -7.6988 3.011e-14
## year_of_observation1988          -2.0834e-01  2.6309e-02  -7.9189 5.747e-15
## year_of_observation1989          -2.7913e-01  2.8052e-02  -9.9504 < 2.2e-16
## year_of_observation1990          -2.8981e-01  2.9157e-02  -9.9396 < 2.2e-16
## year_of_observation1991          -3.2864e-01  2.9889e-02 -10.9951 < 2.2e-16
## year_of_observation1992          -3.8555e-01  3.0841e-02 -12.5012 < 2.2e-16
## year_of_observation1993          -4.0244e-01  3.1365e-02 -12.8306 < 2.2e-16
## year_of_observation1994          -4.3251e-01  3.2170e-02 -13.4445 < 2.2e-16
## year_of_observation1995          -4.2672e-01  3.3154e-02 -12.8709 < 2.2e-16
## year_of_observation1996          -4.6986e-01  3.4885e-02 -13.4687 < 2.2e-16
## year_of_observation1997          -4.8765e-01  3.5896e-02 -13.5852 < 2.2e-16
## year_of_observation1998          -5.3292e-01  3.6562e-02 -14.5760 < 2.2e-16
## year_of_observation1999          -5.4820e-01  3.6948e-02 -14.8373 < 2.2e-16
## year_of_observation2000          -5.7709e-01  3.7457e-02 -15.4066 < 2.2e-16
## year_of_observation2001          -5.6546e-01  3.7866e-02 -14.9333 < 2.2e-16
## year_of_observation2002          -5.3400e-01  3.8329e-02 -13.9319 < 2.2e-16
## year_of_observation2003          -5.4262e-01  3.8326e-02 -14.1581 < 2.2e-16
## year_of_observation2004          -5.6944e-01  3.9423e-02 -14.4446 < 2.2e-16
## blood_alcohol_limit0.08          -3.0202e-02  1.3328e-02  -2.2660  0.023640
## blood_alcohol_limit0.1           -2.5529e-02  9.2215e-03  -2.7685  0.005725
## administrative_license_revocation -5.9539e-02  1.0267e-02  -5.7991 8.667e-09
## seatbelt1                        -4.6644e-02  1.5078e-02  -3.0936  0.002027
## seatbelt2                         8.0347e-05  1.1098e-02   0.0072  0.994225
## sl70plus                          5.8957e-02  1.1848e-02   4.9760 7.512e-07
## graduated_drivers_license_law    -1.4818e-02  1.2828e-02  -1.1551  0.248277
## percent_population_aged_14to24    2.0532e-02  4.1842e-03   4.9070 1.062e-06
## unemployment_rate                -2.8995e-02  2.6654e-03 -10.8782 < 2.2e-16
## vehicle_miles_driven_per_capita   6.1163e-05  4.8638e-06  12.5752 < 2.2e-16
##
## year_of_observation1981          **
## year_of_observation1982          ***
## year_of_observation1983          ***
## year_of_observation1984          ***
## year_of_observation1985          ***
## year_of_observation1986          ***
## year_of_observation1987          ***
## year_of_observation1988          ***
## year_of_observation1989          ***
## year_of_observation1990          ***
## year_of_observation1991          ***
## year_of_observation1992          ***
## year_of_observation1993          ***
## year_of_observation1994          ***
## year_of_observation1995          ***
## year_of_observation1996          ***
## year_of_observation1997          ***
## year_of_observation1998          ***
## year_of_observation1999          ***
## year_of_observation2000          ***
## year_of_observation2001          ***
## year_of_observation2002          ***
## year_of_observation2003          ***
## year_of_observation2004          ***
## blood_alcohol_limit0.08          *
```

```
## blood_alcohol_limit0.1            **
## administrative_license_revocation ***
## seatbelt1                         **
## seatbelt2
## sl70plus                          ***
## graduated_drivers_license_law
## percent_population_aged_14to24    ***
## unemployment_rate                 ***
## vehicle_miles_driven_per_capita   ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    31.924
## Residual Sum of Squares: 8.7807
## R-Squared:      0.72495
## Adj. R-Squared: 0.70502
## F-statistic: 86.668 on 34 and 1118 DF, p-value: < 2.22e-16
```

```
# Random Model

re.model <- plm(total_fatalities_rate ~   year_of_observation + blood_alcohol_limit +   administrative_l
                    data=data2, index=c("state","year_of_observation"), model="random")

#summary(re.model)
```