Lab 3: Panel Models

US Traffic Fatalities: 1980 - 2004

Contents

1 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?"**
>
> In order to answer this research question, we analyzed the 1980-2004 U.S. traffic fatalities panel dataset for 48 states. We conducted comprehensive exploratory data analysis and visually observed the relationship of the potential predictors with traffic fatalities. We also developed several models to estimate the effect of the potential predictors. Based on our study, we believe that changes in traffic laws affect traffic fatalities rate. After evaluating linear models, fixed effects model and random effects model, we believe that the fixed effects model is the most appropriate model for this dataset. See details in the sections below.

2 (30 points, total) Build and Describe the Data

1. (5 points) Load the data and produce useful features.

   We performed comprehensive exploratory data analysis of this dataset. Please see Appendix 1.1 for the code details. We summarized the key notes about our data processing work as follows:

   1) We created new variables and renamed variables using sensibel variable names.

   2) Some rows of the traffice speed columns (sl55/sl65/sl70/s75/slone) have percentage values. These speed columns are categorical variables and should reflect binary values (0, 1). We classified the percentage values based on the field with the max value.

3) The speed_limit_70plus column has values that are not 0 or 1. We reclassified these values to 0 or 1, based on speed_limit column. We also reclassified the states with no speed limit as 1 in this column.

4) We observed non-binary (percentage) values in the three law columns (zero_tolerance_law, graduated_drivers_license_law, per_se_laws). Since we expect these columns to have binary values (0,1), we reclassified all non-zero values as 1 to reflect these fields as binary variables. The minimum_drinking_age column has values that are not integers and we rounded these values to the nearest integer.

5) The *speed_limit* is not set for State 27 between 1996 to 2004. Based on our research, we noted that: for three years after the 1995 repeal of the increased 65 mph limit, Montana had a non-numeric "reasonable and prudent" speed limit during the daytime on most rural roads. But it doesn't mean there was no speed limit. We decided to set the *speed_limit* to 85 for Montana between 1996 to 1999, given the legal case of State v. Rudy Stanko (1998), who got charged for speed of 85 mph. Effective May 28, 1999, as a result of that decision, the Montana Legislature established a speed limit of 75 mph. So we set the *speed_limit* to 75 for Montana between 2000 to 2004.

2. (5 points) Provide a description of the basic structure of the dataset.

   This data set is a balanced longitudinal dataset and contains traffic fatalities data for the 48 continental U.S. states from 1980 through 2004. For each year of observation, the dataset contains state-level cross sectional measurements of fatality count and rate. This data is collected and distributed by Jeffrey M. Wooldridge through this link. In this dataset, the total_fatalities_rate is defined as total fatalities per 100,000 population.

   Wooldridge sourced the data from the National Highway Traffic Safety Administration (NHTSA). Since the early 1980s, NHTSA has been obtaining, from various states, computer data files coded from police accident reports. NHTSA maitains these data files at its National Center for Statistics and Analysis (NCSA). NCSA's goal is to include all states in the program over time, providing a complete census of national traffic statistics. See this link

   After our data processing work, the clean dataset has 25 columns, see Appendix 1.2. These fields include: 1) index variables (year_of_observation, state, year); 2) nine fatality variables: There are three measurements (fatality count, fatality count per 100M miles and fatality rate). Each measurement is provided for total, nighttime and weekend; 3) eight traffic laws indicators (seatbelt, zero_tolerance_law, graduated_drivers_license_law, per_se_laws, minimum_drinking_age, speed_limit_70plus, speed_limit, blood_alcohol_limit), two driving variables (vehicle_miles, vehicle_miles_per_capita), and three economics and demographic variables (statepop, unemployment_rate, pct_population_14_to_24).

3. (20 points) Conduct a very thorough EDA.

   In this dataset, the total_fatalities_rate is defined as total fatalities per 100,000 population.

   Figure 1 below shows the average of total_fatalities_rate by year, with a declining trend visually.

   Next we examined the histograms of the continuous variables and noted that most of these variables are right-skewed. Data transformation (e.g. log) can be helpful.

   We also examined the distribution of the categorical variables (see Appendix 1.3). We observed: 1) Blood alcohol limit: Most states have the limit of 0.1; 2)Minimum drinking age: Most states have 21; 3) Graduated drivers license law: Most states have 0;4) Speed limit: Most states have less than 70 miles; 5) per-se_laws, seatbelt and zero_tolence_law are more evenly distributed among their categories.

   We plotted total_fatalities_rate by state below to understand the fixed effects by state (see Appendix 1.4). We observed strong differences in total traffic fatalities rates across states, suggesting that state-level fixed effects are important for controlling for unobserved differences.

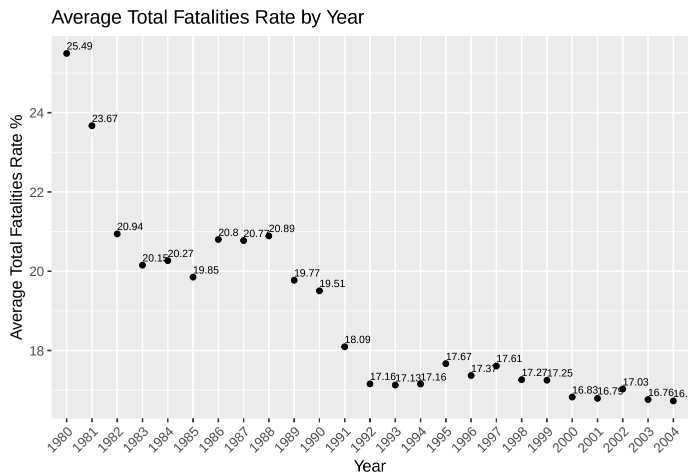


Figure 1: Average Total Fatalities Rates 1980-2004

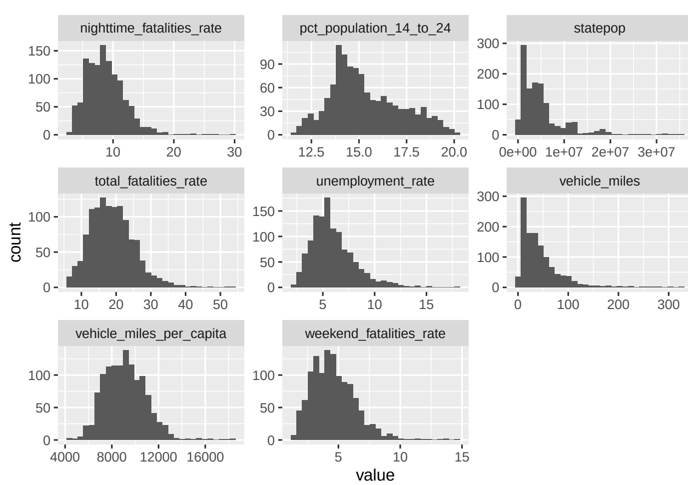


Figure 2: Continuous Explanatory Variables Distribution

We also plotted total_fatalities_rate by state in Figure 3 to understand the trend over the years. All plots are sequentially ordered, despite of missing state number 2, 9, and 12. Most states have a downward trend in total fatalities rate over the years, with trend variations by state.

Figure 3: Total Fatality Rate by State

Figure 4 below compares the original data vs. log-transformed data of several variables: vehicle_miles, vehicle_miles_per_capita, statepop, pct_population_14_to_24, unemployment_rate. The log transformation seemed to improve the relationship between some of these explanatory variables and total fatality rates, as the original data was more skewed. We decided to use log transformation for our interpretation of vehicle_miles, vechicle_miles_per_capita, statepop and unemployment_rate.

Total fatalities rate is positively correlated and percentage of population aged 14-24. Total fatalities rate is negatively correlated with vehicle miles and state population which is not intuitive. We interpret this as population and vehicle miles are both increasing with time, so are other potential factors (e.g. car quality, technology, road conditions, etc.). we noted total fatalities rate is positively correlated with vehicle miles per capita, which is consistent with our background knowledge. Therefore, we will use vehicle miles per capita, instead of vehicle miles and state population separately.

The boxplots (Figure 5) displays the differences in total fatalities rate for the groups within each law categorical variable. We observed that stricter law requirements tend to associate with lower total fatalities rate, see below the detailed comments for each law variable.

1) Seatbelt: No seatbelt requirements tend to have higher fatalities rate than primary and secondary seatbelt requirements; 2) For zero_tolerance_law, graduated_drivers_license_law and per_se_laws: 0 (stricter laws) tends to have lower fatalities rate, comparedto having some level of tolerance in law requirements; 3) minimum_drinking_age: Higher legally-eligible age tends to have lower fatalities rate than lower age limit; 4) speed_limit_70plus: Below 70 speed limit tends to have lower fatalities rate than those with 70+ speed limit; 5) blood_alcohol_limit: 0.08 blood alcohol limit tends to have lower fatalities rate, compared to 0.1 blood alcohol limit.

We examined the correlation matrix (Figure 6) of the continuously variables to understand correlation between variables. State population and vehicle miles are almost perfectly correlated. To avoid perfect colinearity (and the reasons stated above in Figure 4), we will only use

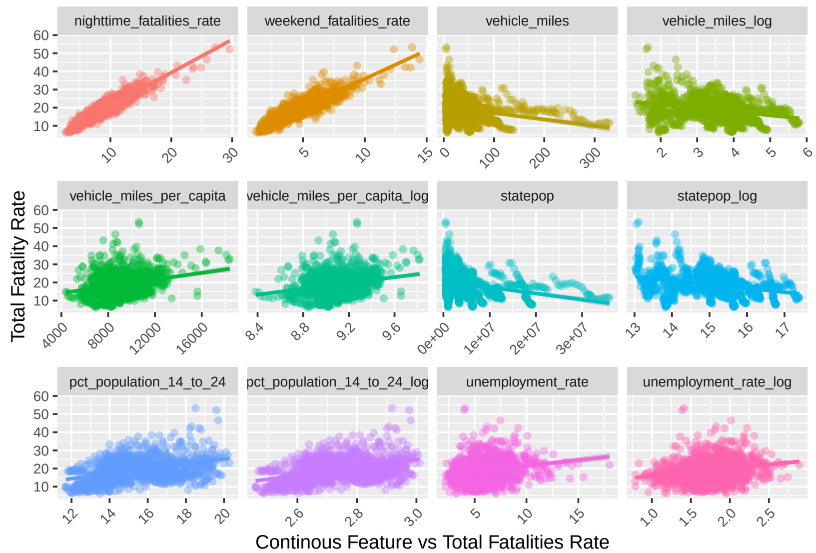


Figure 4: Continous Explanatory Variables Effect on Total Fatality Rate

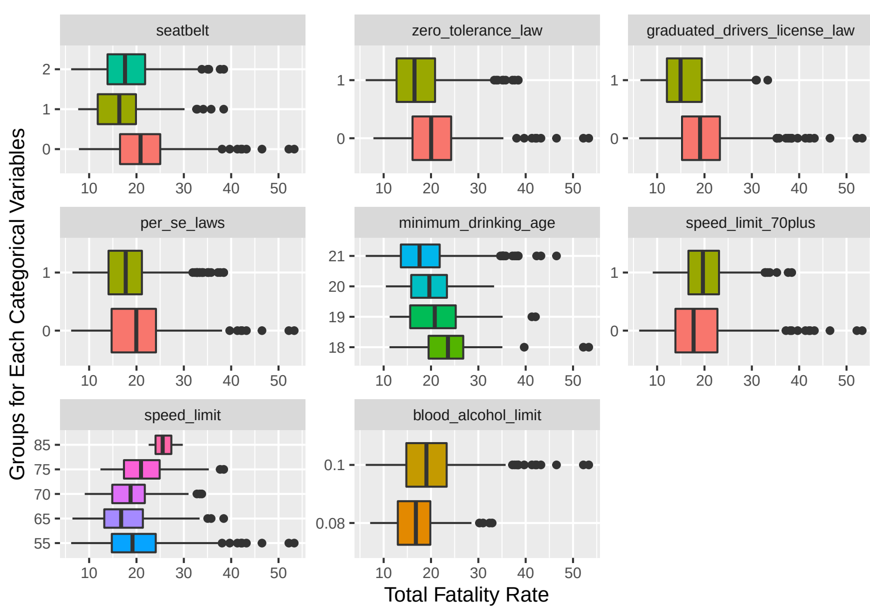


Figure 5: Categorical Variables' Effect on Total Fatalities Rate

vehicle_miles_per_capita in our model development.

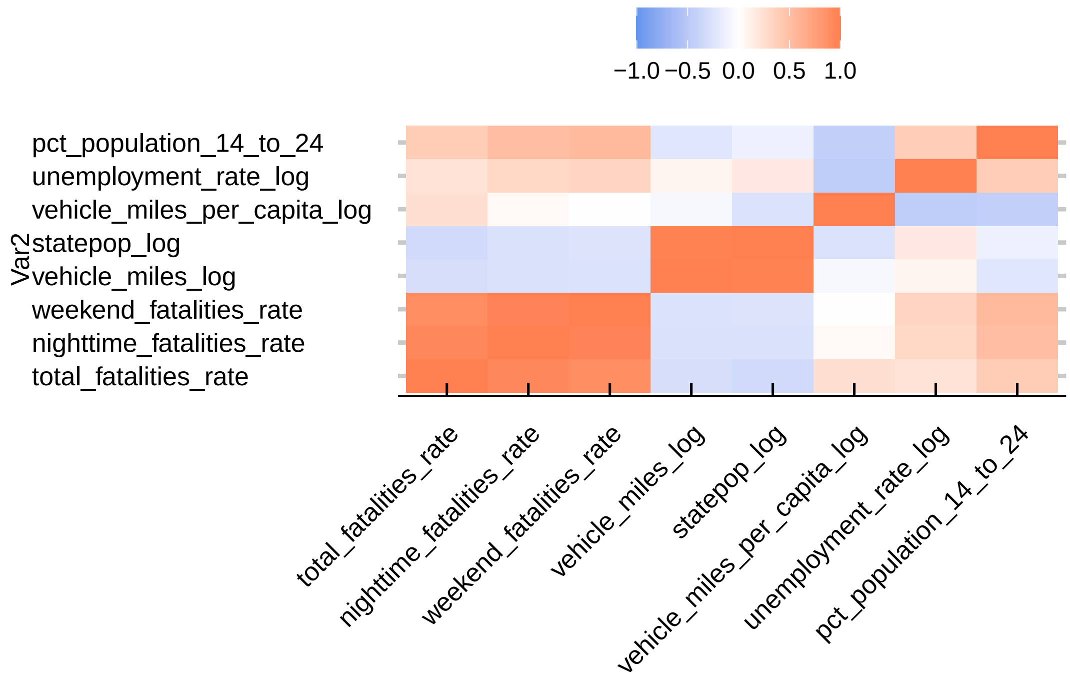


Figure 6: Pearson Correlation

To summarize our data exploratory work, we noted the following key takeaways:

1) We pre-processed the variables in the original dataset, including creating required new variables, cleaning up categorical variables for consistency and renaming all variables with sensible names. Our final dataset in the df file have 25 variables, including index variables (year/year-of-observations and state), the fatality rates variables and the potential explanatory variables (law, economics and demographic variables). We noted 48 states are included in this dataset, which is consistent with the data lable (48 continental states).

2) We conducted comprehensive data exploratory work by using extensive plots to visualize and examine the distribution of each variable, relationship between the total fatalities rate and the potential explanatory variables (including the log form of several variables), as well as the correlation among the variables.

3) We noted a declining trend of the average of total fatalities rates over the years, with state-level trend variations. However, year/time by itself does not fully explain fatalities rates. There are many other factors that affect the total fatalities rates over the years. We had three key observations that could explain the changes in total fatalities rates over the period: a) stricter traffic laws (as indicated in seatbelt, zero_tolerance_law, graduated_drivers_license_law, per_se_laws, minimum_drinking_age, speed_limit_70plus, blood_alcohol_limit) tend to associate with lower total fatalities rate. b) there is a sensible relationship between total fatalities rate and each of the three economics/demographic variables (pct_population_14_to_24, unemployment_rate_log and vehicle_miles_per_capita_log). and finally c) the improvement of technology and road conditions over time can cause lower fatalities. We believe that these potential explanatory variables should be included in the subsequent modeling work.

3 (15 points) Preliminary Model

As a start, fitting a linear model helps identify significant explanatory variables and evaluate whether the linear relationship exists and how strong the linear relationship is. See Appendix 1.5 the "Prelim" column for the model summary results.

```
mod.lm1 <- lm((total_fatalities_rate) ~ year_of_observation, data = df)
```

```
coeftest(mod.lm1, vcov. = vcovHC, type = "HC1")
```

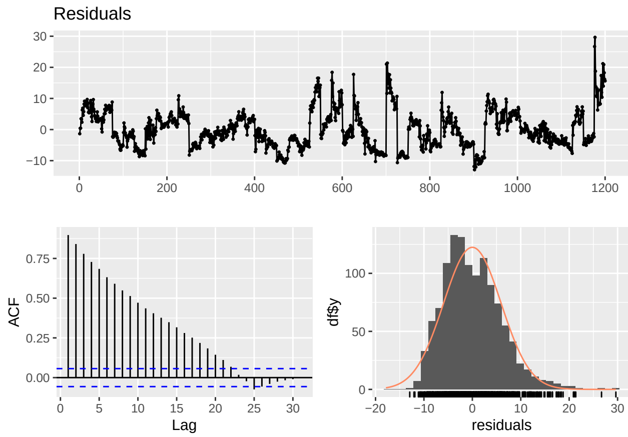


Figure 7: Linear Model Residuals

```
  Breusch-Godfrey test for serial correlation of order up to 28

data:  Residuals
LM test = 1009.4, df = 28, p-value < 2.2e-16
```

We expected the total fatalities rate to decline over the years given our EDA observations. This model shows whether a given year has a linear relationship with total fatalities rate, with 1980 as the base year in the linear model. There is strong statistical evidence that all the years (except for 1981) are negatively related to total fatalities rate at the significance level of 0. The negative relationship aligns with the decreasing trend observed in EDA. Year 1981 is not a significant variable in this model.

As noted in our EDA work, the average total fatalities rate was 24% in 1980 and by 2004 the average total fatalities decreased down to 16%. This linear model shows statistically significant relationship between time and total fatalities rates. However, based on our expericence, time by itself does not fully explain the changes in the fatalities rates. Many other factors (e.g. changes in law, economics and demographics, and technology) could explain the changes in total fatalities rates over the period, as noted in the summary notes of the above EDA section. This dataset is structured by state and by year with many other factors. This introduces omitted variable bias.

Model residuals have large fluctuations for each year and are not constant with zero mean expectation. This violates one of the key classical linear model assumptions - linear conditional mean. The parameter estimates are biased and not reliable. The distribution of residuals is not normal. variance is not constant (heteroskedasticity issues). The Breusch-Godfrey test for serial correlation has a p-value much less than 0.05, suggesting serial correlation. The ACF plot shows significant positive autocorrelation of many lags. With positive serial correlation, the standard errors/uncertainty estimates are understated and the statistical inferences are not reliable. Because of this, we used heteroskedasticity-robust standard errors which are notably higher than the normal standard errors.

4 (15 points) Expanded Model

In the EDA section, we processed data and examined the histograms of continuous variables. We noted right-skewness in the distribution and applied the log transformation to total_fatalities_rate, unemployment_rate, and vehicle miles per capita to normalize the distribution.

In the summary notes of our EDA section, we identified the key variables that can explain total fatalities rate, which align with the variables that we will use in the expanded model. See the model results in Table 1 and Appendix 1.5 in the sections below.

```
mod.lm2 <- lm(
  log(total_fatalities_rate) ~ year_of_observation
    + factor(blood_alcohol_limit)
    + factor(per_se_laws)
    + factor(seatbelt)
    + factor(speed_limit_70plus)
    + factor(graduated_drivers_license_law)
    + pct_population_14_to_24
    + unemployment_rate_log
    + vehicle_miles_per_capita_log,
  data = df
)
```

Parameter estimates for all years are negative, similar to the preliminary model. However, a few added variables have unintuitive signs and are not sensible, for example: seatbelt 2 (secondary seatbelt) has a positive coefficient compared to no seatbelt (even though unsignificant).

Figure 8 shows the model residuals analysis.

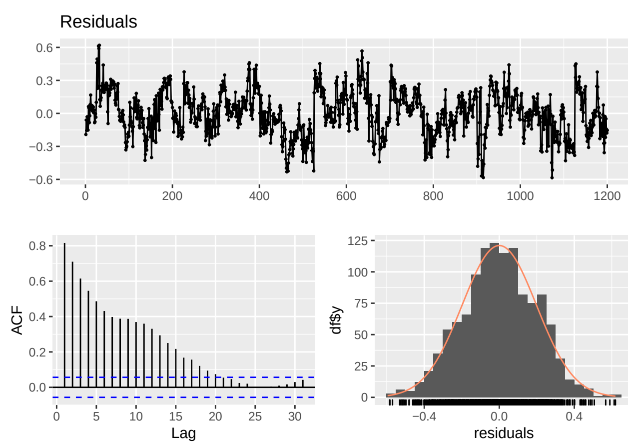


Figure 8: Expanded Linear Model Residuals

```
	Breusch-Godfrey test for serial correlation of order up to 37

data:  Residuals
LM test = 833.32, df = 37, p-value < 2.2e-16
```

The model residuals improved over the prelimary linear model, but still have a lot of variance fluctuations. Breusch-Godfrey test and the ACF plot both indicated serial correlation issues. Standard error estimates and statistical inferences are not reliable. We need to use robust standard errors.

```r
# use the robust standard errors
coeftest(mod.lm2, vcov. = vcovHC, type = "HC1")
```

We evaluated three law variables below.

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

  Every state in the United States has laws making it illegal to operate a motor vehicle while under the influence of alcohol. In each state, the blood alcohol content (BAC) has a legal limit of 0.08% or 0.1% for ordinary, non-commercial vehicles.

  blood_alcohol_limit of 0.1 is the base value in this model. The estimated coefficient on *blood alcohol limit 0.08* is -0.045 and significantly different from zero at 5%. The interpretation is: holding other factors constant, changing the legal blood alcohol limit from 0.10% to 0.08% will cause the traffic fatalities rate to be 4.4% lower (or 0.956 times as big as the original fatalities rate). For example, if the fatalities rate with 0.1% blood alcohol limit was 20%, the estimated fatalities rate would be 19.12% with 0.08% blood alcohol limit.

```
factor(blood_alcohol_limit)0.08
                        -0.0444
```

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

  Yes, the estimated coefficient on *per se laws* is -0.022 and aligns with our expectation. However, this coefficient is not significantly different from zero in this model. The interpretation is that holding other factors constant, changing the per se laws from no (0) to yes (1) will cause traffic fatalities rate to decrease by 2.17% (or 0.9783 times as big as the original fatalities rate). However, this effect is not statistically significant.

```
factor(per_se_laws)1
             -0.0217
```

- Does having a primary seat belt law?

  Yes, the estimated coefficient on *primary seatbelt laws* (seatbelt1) is negative (-0.00067) and the negative sign aligns with our expectation. But it is not significantly different from zero in this model. This is a surprise, as we expect primary seatbelt law requiremnets to have a material impact on the total fatalities rate. Its interpretation is that changing the seatbelt law from no requirement to primary requirement will cause traffic fatalities rate to decrease marginally (0.067%), not statistically or practically significant.

```
factor(seatbelt)1
    -0.0006712344
```

5 (15 points) State-Level Fixed Effects

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

We used the same set of variables in the **Expanded Model** and added factor(state) to account for time-invariant fixed effects in the State-level fixed effects model.

```r
# estimate the fixed effects regression with plm()
mod.fe <- plm(
  log(total_fatalities_rate) ~ year_of_observation
    + factor(blood_alcohol_limit)
    + factor(per_se_laws)
    + factor(seatbelt)
    + factor(speed_limit_70plus)
    + factor(graduated_drivers_license_law)
```

```r
    + pct_population_14_to_24
    + unemployment_rate_log
    + vehicle_miles_per_capita_log
    + factor(state),
  data = df,
  index = c("state", "year_of_observation"),
  model = "within",
  effect = "individual"
)
```

```r
coeftest(mod.fe, vcov. = vcovHC, type = "HC1")
```

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

  In this State-level fixed effects model, the estimated coefficient on blood alcohol limit 0.08 is -0.0048883, much less negative than the estimated coefficient of -0.045 in the expanded model.

  Additionally, the expanded linear model (mod.lm2) showed that this variable is statistically significant, but the State-level fixed effects model shows that this variable's coefficient is not significantly different from zero using the robust standard errors.

  Its interpretation is that changing the blood alcohol limit from 0.10% to 0.08% causes traffic fatalities rate to decrease marginally by 0.49%, which is less than the 4.4% decrease estimated by the expanded liniear model.

```
factor(blood_alcohol_limit)0.08
                        -0.0049
```

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

  The estimated coefficient on *per se laws* is -0.055 and significantly different from zero at 0% significance level, using the robust standard errors. In the prior expanded linear model (mod.lm2), this variable has a coefficient of -0.022 and was not statistically significant.

  Its interpretation is that change per se laws from no (0) to yes (1) will cause traffic fatalities rate to decrease by 5.4%. This is significant both statistically and practically. Compared to the expanded model, the impact magnitude of per se laws estimated by the fixed effects model is much larger and is sensible.

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

  The estimated coefficient on *primary seatbelt laws* is -0.041 and has strong evidence that it is difference from zero at 0.05% significance level (using standard errors) but not significant using the robust standard errors. The prior expanded model (mod.lm2) estimated a much less coefficient.

  Its interpretation is that chaning seatbelt law requirement from none to primary will cause traffic fatalities rate to decrease by 4.1%. This effect makes larger than the marginal effect estimated by the expanded model.

Which set of estimates do you think is more reliable? Why do you think this?

> For the State-level Fixed Effects models, the "within" setting eliminated the omitted variable bias for time-invariant fixed effects and improved the model performance over the linear models. The linear models have omitted variable bias which caused the parameter estimates to be biased.
>
> We performed residuals diagnostic analysis for all models. For the State-level Fixed Effects model, the model residuals have a conditional mean around zero, with a normal distribution. Since this

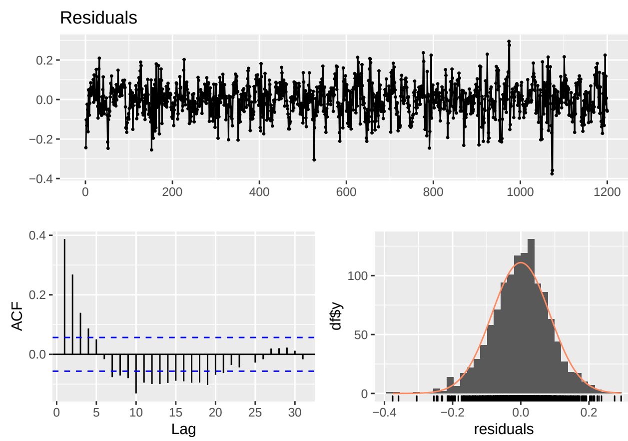


Figure 9: Fixed-Effect Model Residuals

model uses the within setting to fit a fixed effect model at the state level, the effect of unobserved individual heterogeneity is eliminated. The conditinoal mean of zero assumption is met.

Variance fluctulations are much improved compared to the linear models. The ACF plot for the model residuals shows autocorrelation but decays quickly (this is much improved compared to the significant but non-decaying autocorrelation in the expanded model). We used the robust standard errors to account for the group heteroskedasticity and serial correlation by allowing correlation across time in groups.

Based on the residuals analysis, the State-level Fixed Effects model will generate more reliable estimates than the linear models. Also, the parameter estimates' sign and magnitude in the State-level Fixed Effects model make much sense than the expanded model, based on our EDA work and our experience.

- What assumptions are needed in each of these models?

  Fixed Effect Model Assumptions:

  1) **Linearity**: the model is linear in parameters; 2) **i.i.d.** : The observations are independent across individual states but not necessarily across time. This is guaranteed through random sampling. 3) **Indentifiability**: the regressors, including a constant, are not perfectly collinear, and all regressors (but the constant) have non-zero variance and not too many extreme values. 4) **Zero conditional means (strict exogeneity)**

- Are these assumptions reasonable in the current context?

  1) **Linearity**: This assumption is met since each of the explanatory variable is shown to have a linear relationship with the response variable in the models.

  2) **i.i.d.** : This assumption is met since the cross-sectional variables measured across state/year on a randomized population. Since we have 48 different states, we assume that each state has independent legislative procedures and has the state-specific economics and demographic situations.

  3) **Indentifiability**: In our EDA work, we excluded the variables that have perfect collinearity. Also any variables with perfect collinearity would be dropped out from the model. So this condition is met. Using robust standard error on the fitted model, each explanatory variables resulted in a less than 1 standard deviation, which, upon raising it to the power of two, resulted in low non-zero variance. This suggests there are not too many extreme values in the explanatory variables. Therefore, this condition is met.

4) **Zero conditional means (strict exogeneity)**: Since we're using within estimator to fit a fixed effect model on the state, the effect of unobserved individual heterogeneity is eliminated. Thus this condition is met by the State-level Fixed Effect Model. This condition is not met by the two linear models.

6 (10 points) Consider a Random Effects Model

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

  Random Effect Model Assumptions: All assumptions under Fixed Effect model. Additionally, the unobservbed effect term a_i is independent of all explanatory variables in all the time periods.

- 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.

  The first set of assumption is already met under the fixed effect analsysis. To test for the last assumption, we will perform Hausman Test for Fixed vs. Random Effects. In the null hypothesis, both estimates are consistent but only the random effects model estimates are efficient (minimum variance). The alternative hypothes is that only the coefficients of the fixed effects model are consistent, and the coefficients of the random effects model are not consistent. The alternative hypothesis means that there is correlation between residuals and predictors. If Hausman test p-value is less than the significance level, we reject the null hypothesis and deem that the fixed-effects model should be preferred instead of a random-effects model.

  Our test result below shows a p-value of 2.649e-7, which is significantly lower than the 95% confidence level, alpha = 0.05. Thus, there are sufficient evidence to reject the null hypothesis that a random effects model is appropriate, suggesting that we should use the fixed effect models. The random effects model is not likely to be consistent in this case, which means that the paramater estimates won't get closer to the the true parameter values even with a large sample size. Based on this result, we don't think it is appropriate to use the random effects model for this dataset.

```
re.model <- plm(
  log(total_fatalities_rate) ~ year_of_observation
    + factor(blood_alcohol_limit)
    + factor(per_se_laws)
    + factor(seatbelt)
    + factor(speed_limit_70plus)
    + factor(graduated_drivers_license_law)
    + pct_population_14_to_24
    + unemployment_rate_log
    + vehicle_miles_per_capita_log,
  data = df,
  index = c("state", "year_of_observation"),
  model = "random"
)
# Hausman test below shows a p-value of 2.649e-7
phtest(mod.fe, re.model)
# show partial coefficient of state-level time-invariant fixed effects
ranef(re.model)
```

While we don't think it is appropriate to use the random effect model, we included the random effect model in the comparison below for information only. We decided to remove the date effect here to save spaces, please check Appendix 5.1 for the full models including the dates

Table 1: Estimated Models on Total Fatality Rate - Exclude Year Variables (See Appendix 1.5 for full list)

	Dependent variable:			
	(total_fatalities_rate)	log(total_fatalities_rate)		
	OLS	*OLS*	*panel linear*	
	Prelim	Expand	Fixed Effects	Random Effects
	(1)	(2)	(3)	(4)
factor(blood_alcohol_limit)0.08		-0.045^{**} (0.018)	-0.005 (0.017)	-0.006 (0.011)
factor(per_se_laws)1		-0.022 (0.016)	-0.055^{***} (0.017)	-0.053^{***} (0.010)
factor(seatbelt)2		0.019 (0.016)	0.005 (0.017)	0.005 (0.011)
factor(seatbelt)1		-0.001 (0.025)	-0.041 (0.025)	-0.039^{***} (0.015)
factor(speed_limit_70plus)1		0.221^{***} (0.022)	0.073^{***} (0.022)	0.076^{***} (0.012)
factor(graduated_drivers_license_law)1		-0.034^{*} (0.020)	-0.031 (0.020)	-0.031^{**} (0.013)
pct_population_14_to_24		0.018^{*} (0.011)	0.019^{*} (0.011)	0.020^{***} (0.004)
unemployment_rate_log		0.267^{***} (0.024)	-0.194^{***} (0.024)	-0.175^{***} (0.017)
vehicle_miles_per_capita_log		1.541^{***} (0.137)	0.668^{***} (0.130)	0.754^{***} (0.050)
Constant	25.495^{***} (0.438)	-11.304		-3.529^{***} (0.453)
Observations	1,200	1,200	1,200	1,200
R^2	0.128	0.668	0.729	0.712
Residual Std. Error	6.008 (df = 1175)	0.201 (df = 1166)		
Note:				*p<0.1; **p<0.05; ***p<0.01

- 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?

  Since the null hypothesis for the Hausman test is rejected, the last assumption is not met. Thus we should not use the random effect model to estimate the coefficients since the estimates will not likely to be consistent.

  The random effects model assumes the individual unobserved heterogeneity is uncorrelated with the independent variables. However, the Hausman test shows strong evidence that correlation exists. Therefore, using the random effects model for this dataset could produce bias in parameters estimates. This bias does not arise in the fixed effects model, because the fixed effects model assumes that the time-invariant fixed effects can be correlated with the other variables. Fixed effects models remove the effect of time-invariant variables so we can assess the net effect of the other predictors on the response variable (total traffic fatalities rate).

  Given the Hausman test result, the random effects model is also not likely to be efficient, which means its variance would not be minimum or stable causing the standard errors to be biased and test statistics to be not reliable.

  From a practical perspective, fixed effects models are relatively straighforward. Random effects models require additional mathematical assumptions with added-complexity, which strenthens the argument of not using the random effects model in this case.

  This is especially true because there is no time invariant explanatory variables in our model, so the biggest advantage of the random effects model (being able to estimate the impacts of time-invariant explanatory variables) is not applicable here.

7 (10 points) Model Forecasts

We collected the United States Motor Vehicle Miles Traveled Total (Millions) data from the US Department of Transformation from Jan 1980 to January 2023. This data is available at here [https://www.fhwa.dot.gov/policyinformation/statistics/2021/vm202.cfm]. We also downloaded related Bloomberg data (under ticker: VMTDVCLE Index) and saved it in the file `data/VMTDVCLE.csv`.

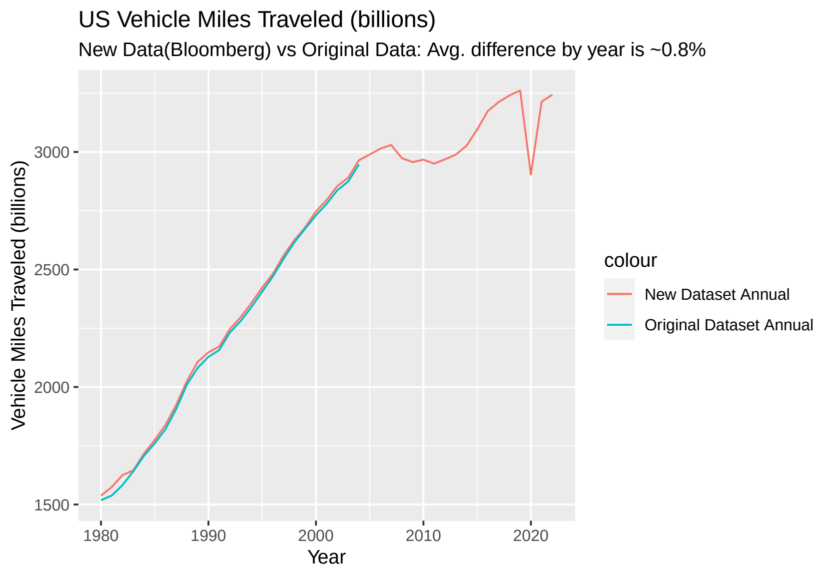


Figure 10: Comparison of Origianl and New Vehicle Miles

Figure 6 shows that new vehicle milage data is similar to the old data. The average difference between the two data is marginal (0.8%). This difference is expected as the new vehicle miles data from Bloomberg includes 51 states, while the old data only includes 48 states. We believe that the new data is appropriate to use in our analysis.

The largest decrease in driving during COVID bust is in April 2020 (-39.1% compared with April 2018), and the largest increase in driving post-COVID boom is in September 2022 (+5.0% compared with January 2018). We didn't consider January and February in 2020 because COVID hasn't started yet then.

month	2019	2020	2021	2022
1	1.7	6.6	-8.1	-4.3
2	1.8	6.6	-6.5	3.5
3	0.6	-16.3	-0.5	2.5
4	-0.6	-39.1	-5.8	-4.4
5	2.1	-22.1	0.2	1.5
6	-0.5	-11.4	1.5	-0.2
7	0.2	-8.7	1.9	-1.5
8	2.9	-7.0	0.8	1.5
9	2.2	-3.7	4.0	5.0
10	0.9	-5.3	1.5	1.6
11	-0.1	-8.5	2.8	1.4
12	-3.2	-10.7	-3.5	-5.2

```
beta <- tail(mod.fe$coefficients, 1)
(1 + 0.05)^beta - 1
```

```
  vehicle_miles_per_capita_log
                    0.03312113
```

```r
(1 - 0.391)^beta - 1
```

```
  vehicle_miles_per_capita_log
                    -0.2819463
```

Holding other factors constant, if vehicle miles per capita increases by 5.0%, the traffic fatalities rate would increase by 3.3% (or become 1.033 times as large), compared with before COVID levels.

Holding other factors constant, if vehicle miles per capita decreases by 39.1%, the traffic fatalities rate would decrease by 28.19% (or become 71.81% as large), compared with before COVID levels. Althought we should be cautious here: the change in miles driven per capita is very large, so the log transformation interpretation might not generate a good estimation in this case.

8 (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?

```r
pcdtest(mod.fe, test = "lm")
```

```r
pbgtest(mod.fe, order = 2)
```

If there is serial correlation or heteroskedasticity in the idiosyncratic errors of the model, the standard errors or uncertainty of the estimators will be underestimated and statistical inferences are not reliable. However, serial correlation does not cause bias in the regression coefficient estimates of the estimators.

In this case, when we apply the Breusch Pagan Test on homoskedasticity to the FE model, we obtained a p-value of 2.2e-16, which is significantly less than 0.05, suggesting that there is sign of heteroskedasticity from the idiosyncratic errors. We also performed the Breusch-Godfrey test for serial correlation and obtained a p-value of 2.2e-16, which is significantly less than 0.05, suggesting that we have do have serially correlation in idiosyncratic errors. In the ACF plot, we can also visually detect the autocorrelation in the model residuals.

Therefore, the FE model has serial correlation issues and standard errors are underestimated. We should use the robust standard errors to address heteroskedasticity and serial correlation issues by allowing correlation across time in groups.

9 Appendix

Appendix 1.1

```
# Appendix 1.1
# produce new variables - year_of_observation, speed_limit, speed_limit_70plus, blood_alcohol_limit
df <- data %>%
  mutate(state = factor(state)) %>%
  rowwise() %>%
  # speed_limit
  mutate(
    speed_limit_70plus = factor(sl70plus),
    speed_limit = parse_number(
      colnames(
        select(data, starts_with("sl"))
      )[which.max(c_across(starts_with("sl")))],
      na = "slnone"
    ),
  ) %>%
  select(-starts_with("sl")) %>%
  mutate(year_of_observation = factor(year)) %>% # year_of_observation
  select(-starts_with("d")) %>%
  mutate(blood_alcohol_limit = factor(parse_number(
    colnames(
      select(data, starts_with("bac"))
    )[which.max(c_across(starts_with("bac")))]
  ) / 100)) %>% # blood_alcohol_limit
  select(-starts_with("bac")) %>%
  mutate(
    seatbelt = factor(seatbelt), # 'seatbelt' categorizes primary or secondary
    speed_limit_70plus = ifelse(speed_limit == 55 | speed_limit == 65, 0, 1)
  ) %>%
  select(-starts_with("sb"))

# rename the variables to sensible names
df <- df %>%
  dplyr::rename(
    "total_fatalities_rate" = "totfatrte",
    "minimum_drinking_age" = "minage",
    "zero_tolerance_law" = "zerotol",
    "graduated_drivers_license_law" = "gdl",
    "per_se_laws" = "perse",
    "total_traffic_fatalities" = "totfat",
    "total_nighttime_fatalities" = "nghtfat",
    "total_weekend_fatalities" = "wkndfat",
    "total_fatalities_per_100_million_miles" = "totfatpvm",
    "nighttime_fatalities_per_100_million_miles" = "nghtfatpvm",
    "weekend_fatalities_per_100_million_miles" = "wkndfatpvm",
    "nighttime_fatalities_rate" = "nghtfatrte",
    "weekend_fatalities_rate" = "wkndfatrte",
    "vehicle_miles" = "vehicmiles",
    "unemployment_rate" = "unem",
    "pct_population_14_to_24" = "perc14_24",
    "vehicle_miles_per_capita" = "vehicmilespc"
  ) %>%
```

```r
  select(
    year_of_observation,
    state,
    year,
    # response variables
    total_fatalities_rate,
    nighttime_fatalities_rate,
    weekend_fatalities_rate,
    total_traffic_fatalities,
    total_nighttime_fatalities,
    total_weekend_fatalities,
    total_fatalities_per_100_million_miles,
    nighttime_fatalities_per_100_million_miles,
    weekend_fatalities_per_100_million_miles,
    # potential explanatory variables
    seatbelt,
    zero_tolerance_law,
    graduated_drivers_license_law,
    per_se_laws,
    minimum_drinking_age,
    speed_limit_70plus,
    speed_limit,
    blood_alcohol_limit,
    vehicle_miles,
    vehicle_miles_per_capita,
    # econ and demographic variables
    statepop,
    unemployment_rate,
    pct_population_14_to_24, vehicle_miles
  ) # keep the similar variables together

# check the data
# df %>% glimpse()
```

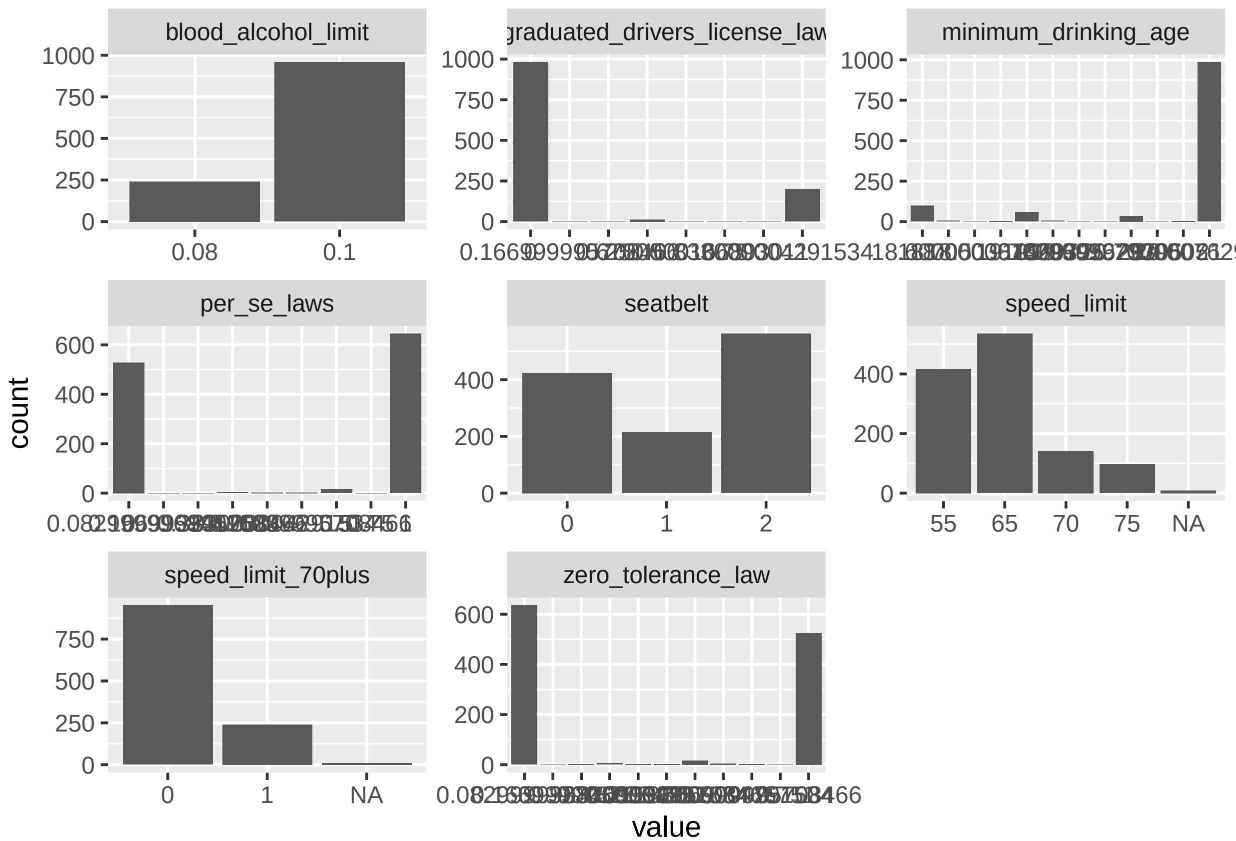
Appendix 1.2

```
# Appendix 1.2
glimpse(df)
```

```
Rows: 1,200
Columns: 25
Rowwise:
$ year_of_observation                      <fct> 1980, 1981, 1982, 1983, 198~
$ state                                    <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
$ year                                     <int> 1980, 1981, 1982, 1983, 198~
$ total_fatalities_rate                    <dbl> 24.14, 24.07, 21.37, 23.64,~
$ nighttime_fatalities_rate                <dbl> 10.84, 11.08, 9.58, 10.09, ~
$ weekend_fatalities_rate                  <dbl> 6.060000, 6.330000, 5.71000~
$ total_traffic_fatalities                 <int> 940, 933, 839, 930, 932, 88~
$ total_nighttime_fatalities               <int> 422, 434, 376, 397, 421, 35~
$ total_weekend_fatalities                 <int> 236, 248, 224, 223, 237, 22~
$ total_fatalities_per_100_million_miles   <dbl> 3.200, 3.350, 2.810, 3.000,~
$ nighttime_fatalities_per_100_million_miles <dbl> 1.437, 1.558, 1.259, 1.281,~
$ weekend_fatalities_per_100_million_miles <dbl> 0.803, 0.890, 0.750, 0.719,~
$ seatbelt                                 <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ zero_tolerance_law                       <dbl> 0.000, 0.000, 0.000, 0.000,~
$ graduated_drivers_license_law            <dbl> 0.00, 0.00, 0.00, 0.00, 0.0~
$ per_se_laws                              <dbl> 0.000, 0.000, 0.000, 0.000,~
$ minimum_drinking_age                     <dbl> 18, 18, 18, 18, 18, 20, 21,~
$ speed_limit_70plus                       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ speed_limit                              <dbl> 55, 55, 55, 55, 55, 55, 55,~
$ blood_alcohol_limit                      <fct> 0.1, 0.1, 0.1, 0.1, 0.1, 0.~
$ vehicle_miles                            <dbl> 29.37500, 27.85200, 29.8576~
$ vehicle_miles_per_capita                 <dbl> 7543.874, 7107.785, 7606.62~
$ statepop                                 <int> 3893888, 3918520, 3925218, ~
$ unemployment_rate                        <dbl> 8.8, 10.7, 14.4, 13.7, 11.1~
$ pct_population_14_to_24                  <dbl> 18.9, 18.7, 18.4, 18.0, 17.~
```

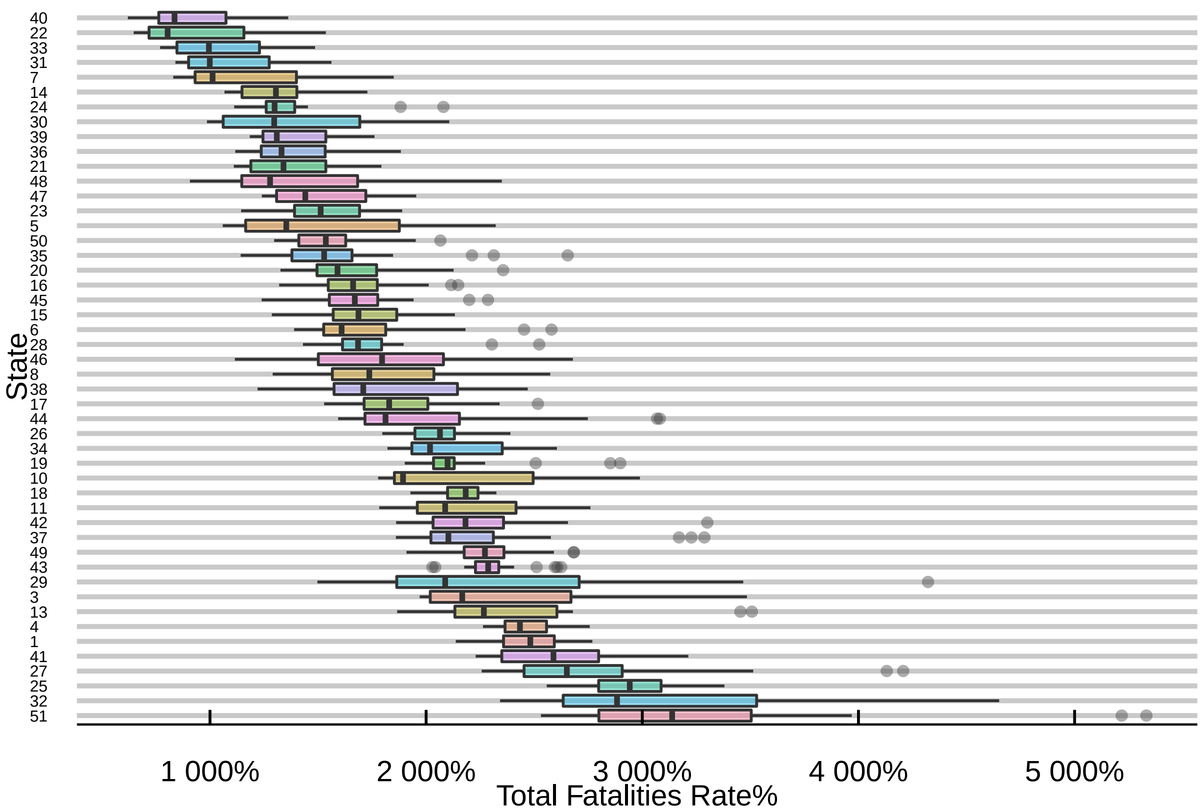
Appendix 1.3

```r
# Appendix 1.3
df %>%
  select(
    seatbelt,
    zero_tolerance_law,
    graduated_drivers_license_law,
    per_se_laws,
    minimum_drinking_age,
    speed_limit_70plus,
    speed_limit,
    blood_alcohol_limit
  ) %>%
  gather() %>%
  ggplot(aes(value)) +
  facet_wrap(~key, scales = "free") +
  geom_bar()
```



> **Appendix 1.4**

```r
# Appendix 1.4
df %>%
  ggplot(aes(reorder(state, desc(total_fatalities_rate)), total_fatalities_rate,
    fill = state
  )) +
  geom_boxplot(alpha = 0.4) +
  theme_economist_white(gray_bg = FALSE) +
  theme(legend.position = "none", axis.text.y = element_text(size = 6)) +
  scale_y_continuous(label = percent) +
  xlab("State") +
  ylab("Total Fatalities Rate%") +
  coord_flip()
```



> **Appendix 1.5**

Table 2: Estimated Models on Total Fatality Rate

	Dependent variable:			
	(total_fatalities_rate)	log(total_fatalities_rate)		
	OLS	*OLS*	*panel linear*	
	Prelim	Expand	Fixed Effects	Random Effects
	(1)	(2)	(3)	(4)
year_of_observation1981	−1.824 (1.615)	−0.091** (0.046)	−0.062*** (0.017)	−0.063*** (0.017)
year_of_observation1982	−4.552*** (1.510)	−0.295*** (0.045)	−0.134*** (0.018)	−0.141*** (0.018)
year_of_observation1983	−5.342*** (1.455)	−0.352*** (0.044)	−0.166*** (0.022)	−0.176*** (0.021)
year_of_observation1984	−5.227*** (1.407)	−0.306*** (0.044)	−0.212*** (0.021)	−0.219*** (0.022)
year_of_observation1985	−5.643*** (1.399)	−0.344*** (0.045)	−0.238*** (0.026)	−0.246*** (0.026)
year_of_observation1986	−4.694*** (1.419)	−0.321*** (0.049)	−0.201*** (0.034)	−0.211*** (0.033)
year_of_observation1987	−4.720*** (1.423)	−0.357*** (0.049)	−0.247*** (0.038)	−0.259*** (0.038)
year_of_observation1988	−4.603*** (1.390)	−0.367*** (0.051)	−0.278*** (0.048)	−0.290*** (0.047)
year_of_observation1989	−5.722*** (1.395)	−0.452*** (0.055)	−0.351*** (0.054)	−0.365*** (0.053)
year_of_observation1990	−5.989*** (1.413)	−0.512*** (0.059)	−0.361*** (0.059)	−0.378*** (0.058)
year_of_observation1991	−7.400*** (1.384)	−0.628*** (0.060)	−0.398*** (0.063)	−0.419*** (0.062)
year_of_observation1992	−8.337*** (1.386)	−0.734*** (0.064)	−0.459*** (0.067)	−0.482*** (0.066)
year_of_observation1993	−8.367*** (1.378)	−0.725*** (0.062)	−0.476*** (0.068)	−0.498*** (0.067)
year_of_observation1994	−8.339*** (1.401)	−0.711*** (0.062)	−0.509*** (0.067)	−0.530*** (0.067)
year_of_observation1995	−7.826*** (1.453)	−0.689*** (0.064)	−0.509*** (0.073)	−0.530*** (0.072)
year_of_observation1996	−8.125*** (1.422)	−0.814*** (0.064)	−0.560*** (0.075)	−0.584*** (0.075)
year_of_observation1997	−7.884*** (1.436)	−0.822*** (0.065)	−0.583*** (0.077)	−0.607*** (0.076)
year_of_observation1998	−8.229*** (1.443)	−0.869*** (0.066)	−0.636*** (0.077)	−0.661*** (0.077)
year_of_observation1999	−8.244*** (1.481)	−0.866*** (0.066)	−0.651*** (0.079)	−0.676*** (0.078)
year_of_observation2000	−8.669*** (1.447)	−0.877*** (0.068)	−0.682*** (0.079)	−0.706*** (0.079)
year_of_observation2001	−8.702*** (1.442)	−0.930*** (0.069)	−0.650*** (0.083)	−0.679*** (0.081)
year_of_observation2002	−8.465*** (1.468)	−0.974*** (0.071)	−0.611*** (0.081)	−0.644*** (0.079)
year_of_observation2003	−8.731*** (1.445)	−0.997*** (0.072)	−0.613*** (0.084)	−0.647*** (0.081)
year_of_observation2004	−8.766*** (1.470)	−0.980*** (0.073)	−0.650*** (0.088)	−0.683*** (0.086)
factor(blood_alcohol_limit)0.08		−0.045*** (0.017)	−0.005 (0.018)	−0.006 (0.017)
factor(per_se_laws)1		−0.022 (0.015)	−0.055*** (0.016)	−0.053*** (0.017)
factor(seatbelt)2		0.019 (0.023)	0.005 (0.016)	0.005 (0.017)
factor(seatbelt)1		−0.001 (0.025)	−0.041* (0.025)	−0.039 (0.025)
factor(speed_limit_70plus)1		0.221*** (0.021)	0.073*** (0.022)	0.076*** (0.022)
factor(graduated_drivers_license_law)1		−0.034 (0.025)	−0.031 (0.020)	−0.031 (0.020)
pct_population_14_to_24		0.018*** (0.007)	0.019* (0.011)	0.020* (0.011)
unemployment_rate_log		0.267*** (0.024)	−0.194*** (0.024)	−0.175*** (0.024)
vehicle_miles_per_capita_log		1.541*** (0.048)	0.668*** (0.137)	0.754*** (0.130)
Constant	25.495*** (1.157)	−11.304*** (0.438)		−3.529*** (1.167)
Observations	1,200	1,200	1,200	1,200
R^2	0.128	0.668	0.729	0.712
Residual Std. Error	6.008 (df = 1175)	0.201 (df = 1166)		
Note:				*p<0.1; **p<0.05; ***p<0.01