|  |
| --- |
|  |
| 20-PBD-002 Shraddha P Jain |
| 3951 – Introduction to Econometrics and Finance End Semester Examination |
|  |
|  |
|  |

|  |
| --- |
|  |

**Panel Data Analysis**

1. **Write a note on Panel Data Analysis**

Ans: Panel data is a dataset in which the behavior of entities is observed across time. These entites can be people, households, firms, states, countries, companies, etc.

Eg:



**There are a number of advantages of panel data:**

* **Panel data can model both the common and individual behaviors of groups.**
* **Panel data contains more information, more variability, and more efficiency than pure**[**time series data**](https://www.aptech.com/blog/introduction-to-the-fundamentals-of-time-series-data-and-analysis/)**or cross-sectional data.**
* **Panel data can detect and measure statistical effects that pure time series or cross-sectional data can't.**
* **Panel data can minimize estimation biases that may arise from aggregating groups into a single time series.**

**Panel (data) analysis is a statistical method, widely used in**[**social science**](https://en.wikipedia.org/wiki/Social_science) **and**[**econometrics**](https://en.wikipedia.org/wiki/Econometrics)**to analyze two-dimensional (typically cross sectional and longitudinal)**[**panel data**](https://en.wikipedia.org/wiki/Panel_data)**. The data are usually collected over time and over the same individuals and then a**[**regression**](https://en.wikipedia.org/wiki/Linear_regression)**is run over these two dimensions**

There are various approaches to panel data analysis, which include

1. Fixed effect models
2. Random effect models

Panel data model is represented as:



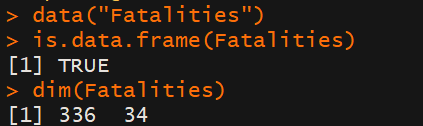
The observable outcome variable of interest is yit. On the right-hand side, we have a constant term, x1it = 1. We include one observable variable, x2it, that has variation across individuals and time. The variable w1i is time-invariant and varies only across individuals. The population parameters β1, β2, and α1 have no subscripts and are fixed in all time periods for all individuals.

We have included only one x-variable and one w-variable to keep things simple, but there can be more of each type.

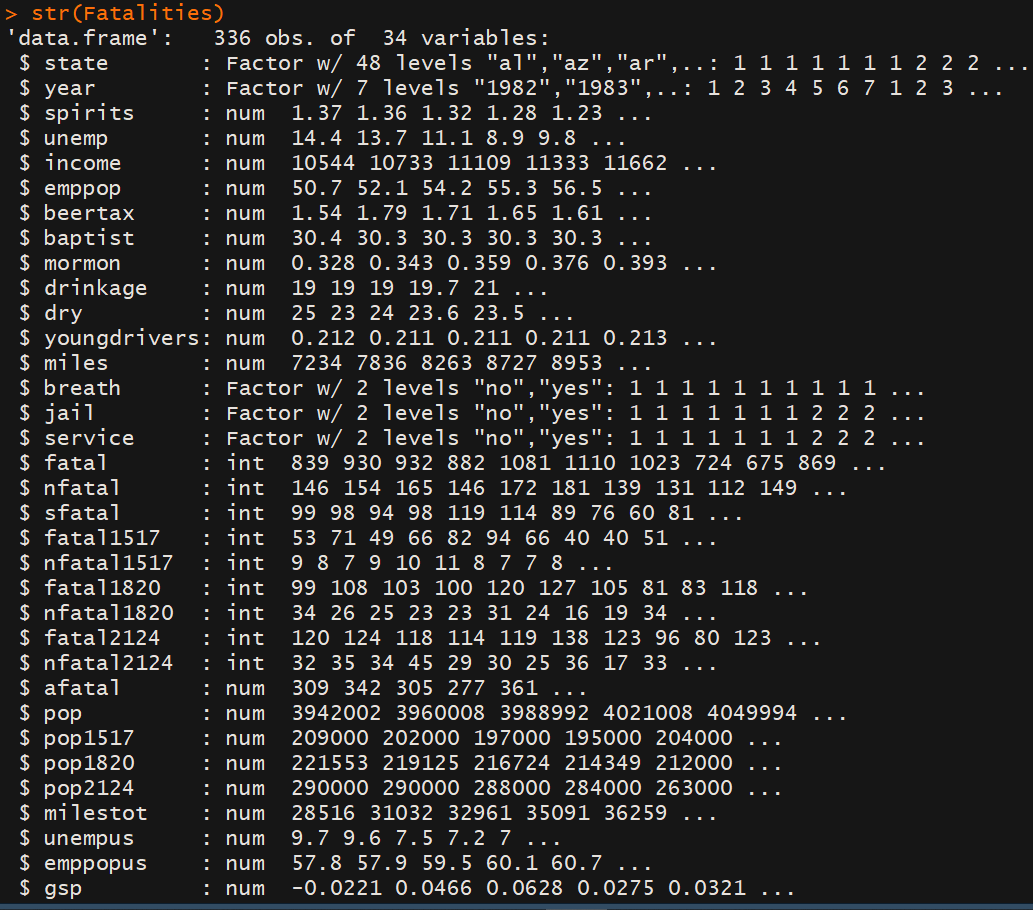
In parentheses, we have the two random error components, one associated with the individual (ui) and one associated with the individual and time (eit).

**2. Use the data ‘fatali4es’ from package AER.**

**Ans:**



The dataset consists of 336 observations on 34 variables.



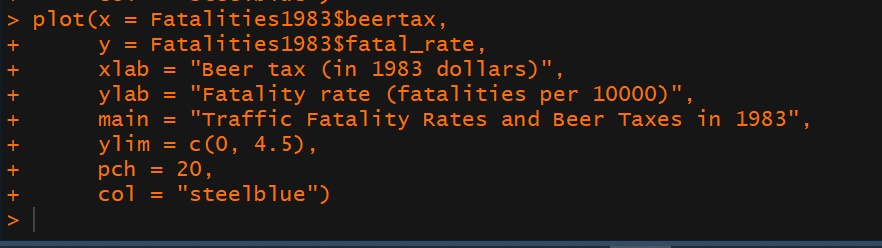
The variable state is a factor variable with 48 levels. (one for each of the 48 contiguous federal states of the U.S.).

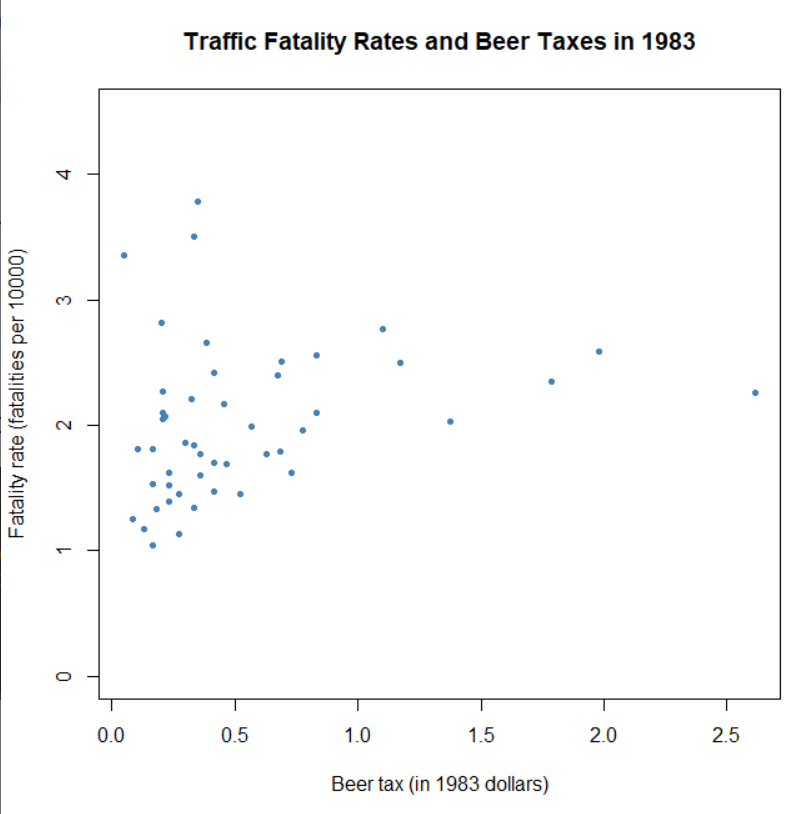
The variable year is also a factor variable that has 7 levels identifying the time period when the observation was made (thus 7 x 48 =336)

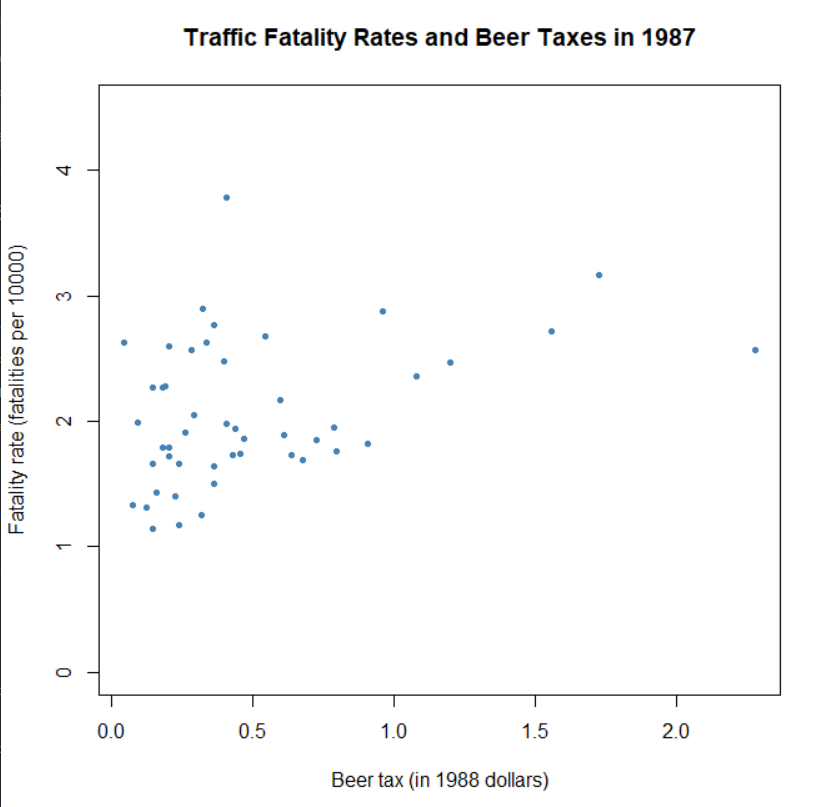
Since all variables are observed for all entities and over all time periods, the panel is balanced.

**3. You need to assess the impact of Alcohol taxes on Traffic Deaths.**

Ans:







In both years 1983, 1987 we see and a somewhat increasing trend of traffic death rates as beer rates increase.

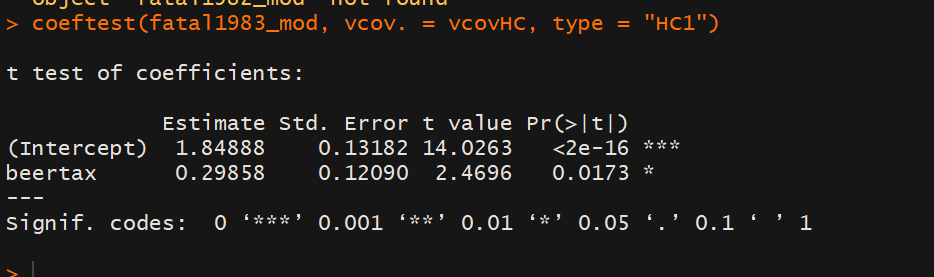
**4. Specify a Simple Linear Regression model for the above**

**Ans.** A simple linear regression model for the above would be:

fatality\_rate\_(year) = β0 + β1\*Beer\_Tax\_(Year) + ϵ

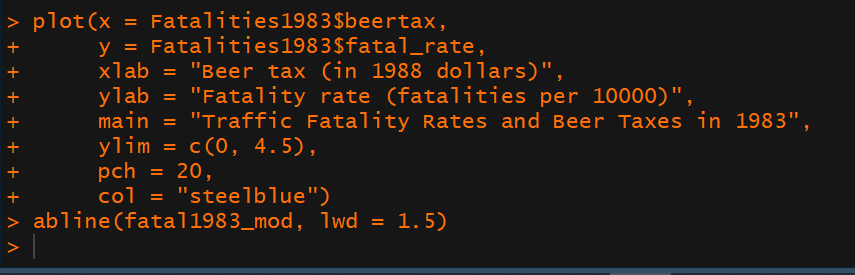
**5. Apply the same in R and interpret the results**

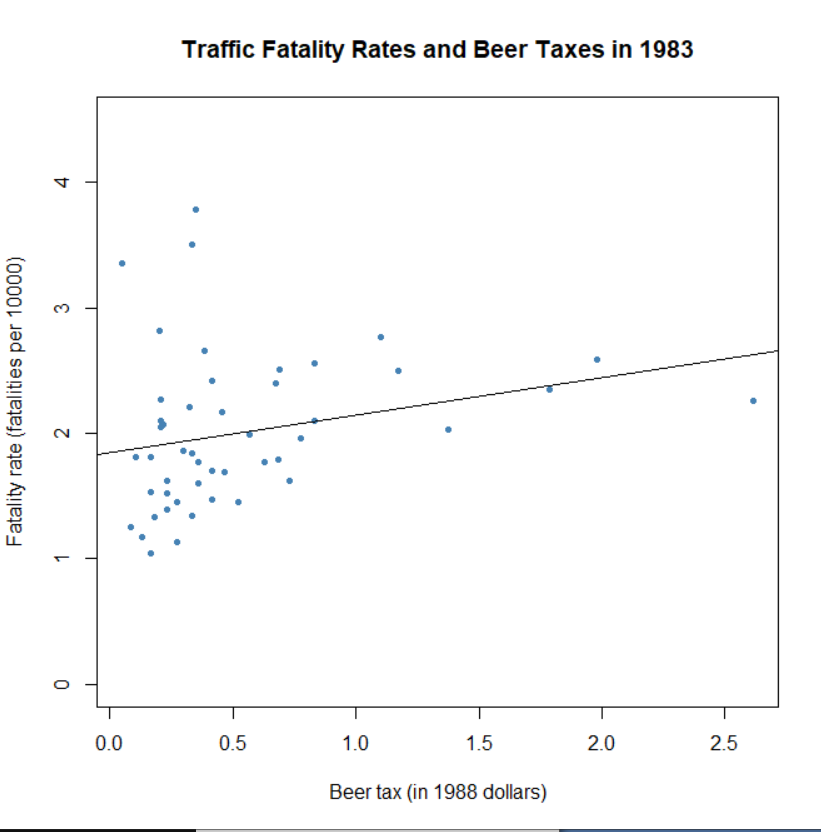
Ans. For the year 1983:



We have the model: fataility\_rate\_1983 = 1.8488+0.29858\*beertax\_1983

The slope and intercept are both significant by the model.





This is the fitted line on the data.

In the plot, each point represents observations of beer tax and fatality rate for a given state in the year 1983. The regression results indicate a positive relationship between the beer tax and the fatality rate for both years. This is contrary to our expectations: alcohol taxes are supposed to *lower* the rate of traffic fatalities. This is possibly due to omitted variable bias, since both models do not include any covariates, e.g., economic conditions.

This could be corrected for using a multiple regression approach. However, this cannot account for omitted *unobservable* factors that differ from state to state but can be assumed to be constant over the observation span.

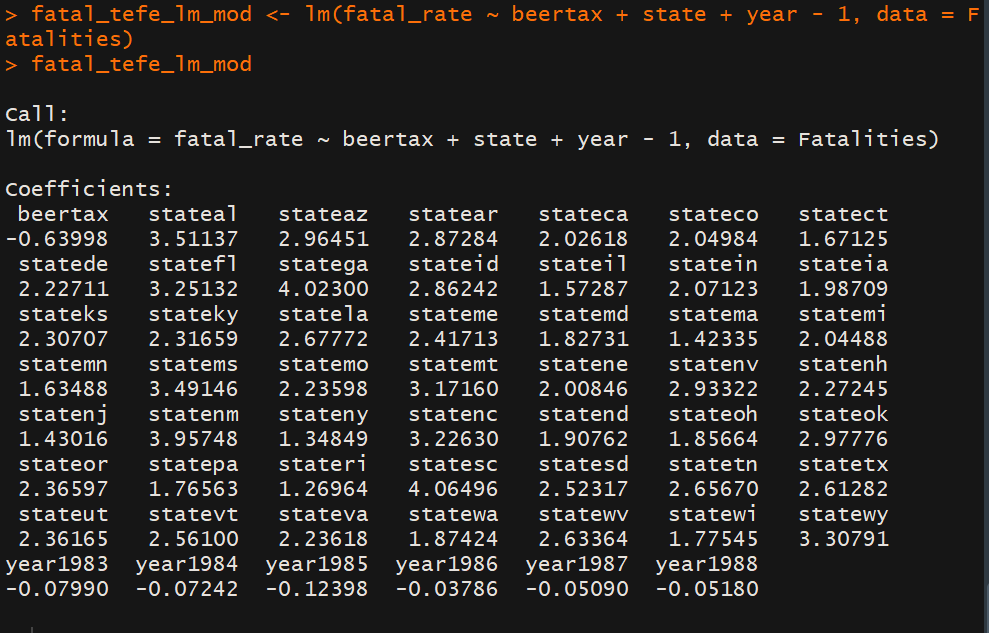
**6. Extend the above model and specify a Panel Data Regression Model for the same research question and explain it.**

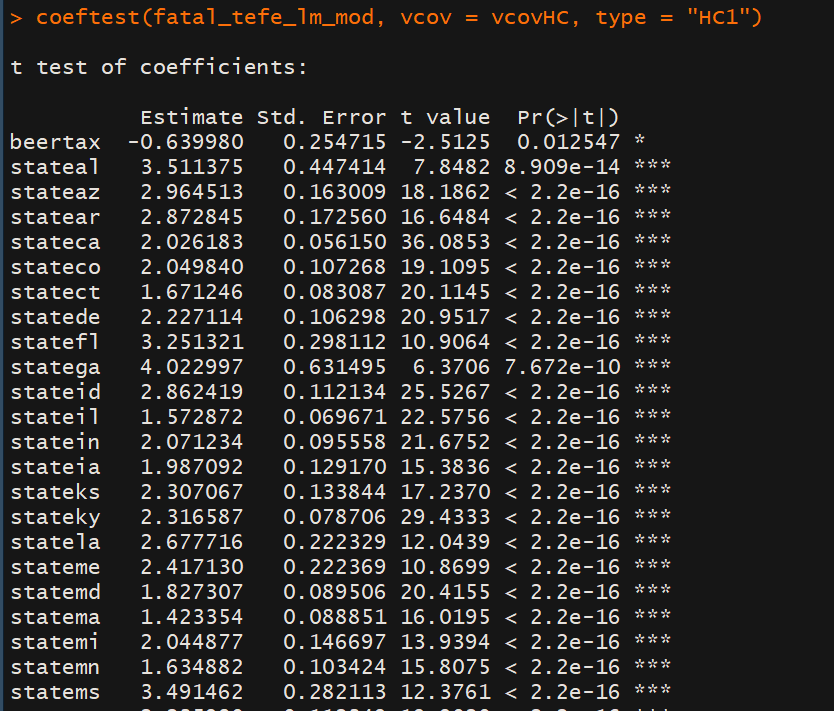
**Ans**. A panel data regression model for the above question would be:

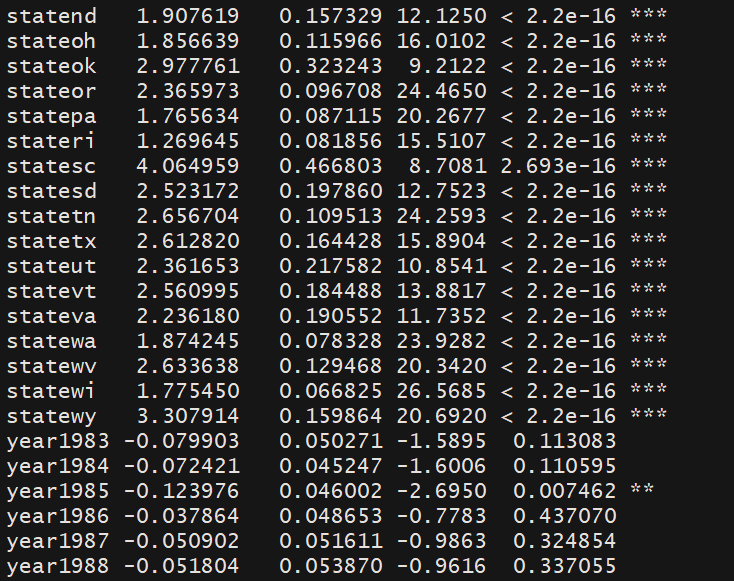
Fatality\_Rateit = β1\*BeerTaxit+ StateEffects+TimeEffects+uit

Fatality Rate is our target variable. BeerTax represents the data that changes across time and individuals. StateEffects variables account for the variables that change across individuals but are constant over time while TimeEffects variables account for the variables that change across time but are constant over individuals. uit represents the error term.

**7. Implement the above in R and interpret the results**

Ans. 





From the above output, we can see that beer taxes, as well as the state effect variables are all significant, while only the effects of year 1985 are significant.

Also, the model we get is of type:

Fatality\_Rateit = -0.6999\*BeerTaxit+ StateEffects+TimeEffects+uit

Here, as we expect, the coefficient of beerTax is negative, signifying that more the bear taxes applied, less is the fatality rate.

**Simultaneous Equation Models**

1. **What do you mean by ILS and 2SLS? Can we use these methods when there is no simultaneity? Justify your answer**

**Ans:**

* in simultaneous-equation models more than one dependent, or endogenous, variable is involved, necessitating as many equations as the number of endogenous variables
* A unique feature of simultaneous-equation models is that the endogenous variable (i.e., regressand) in one equation may appear as an explanatory variable (i.e., regressor) in another equation of the system
* As a consequence, such an endogenous explanatory variable becomes stochastic and is usually correlated with the disturbance term of the equation in which it appears as an explanatory variable
* There are two approaches to estimate the parameters of simultaneous equation models:
  + Indirect Least Squares
  + 2 Stage Least Squares
* Indirect Least Squares Method is used when the system of simultaneous is exactly identified. It has the following steps:
  + Obtain reduced form equations
  + Apply OLS to reduced form equations individually. Estimates are consistent (mainly asymptotically efficient)
  + We obtain estimates of the original structural coefficients from the estimated reduced-form coefficients obtained in Step 2.
  + If an equation is exactly identified, there is a one-to-one correspondence between the structural and reduced-form coefficients; that is, one can derive unique estimates of the former from the latter
* 2 Stage Least Squares Method is used in the case when the system of simultaneous equations are over-identified.
  + The basic idea behind 2SLS is to replace the (stochastic) endogenous explanatory variable by a linear combination of the predetermined variables in the model and use this combination as the explanatory variable in lieu of the original endogenous variable. Thus create a proxy which is not correlated to the error term.
* **No, we cannot use these methods in case of there is no simultaneity in our model. In such a case, we will not have the rank and order conditions satisfied, and will not be able to get the reduced form equations.**

**ARCH/GARCH Models**

1. **Discuss the main idea behind ARCH models**

**Ans. Autoregressive conditional heteroskedasticity (ARCH) is a statistical model used to analyze**[**volatility**](https://www.investopedia.com/terms/v/volatility.asp)**in time series in order to forecast future volatility.**

* ARCH/GARCH models focus on the assumption of Homoskedasticity.
* In the presence of heteroskedasticity, the regression coefficients of OLS are still unbiased, but the standard errors and confidence intervals will be too narrow, giving a false sense of precision.
* The key issue is the variance of the error terms and what makes them large. This question often arises in financial applications
* Even a cursory look at financial data suggests that some time periods are riskier than others; that is, the expected value of the magnitude of error terms at some times is greater than at others.
* The most common models used to capture varying variances are the Autoregressive Conditional Heteroskedasticity (ARCH) model.
* Conditional Variance: conditioned on the entire set of outcomes previous to time t . That is, at time t , we will presume to know the realized values of all past variables
* Advantages:
  + ARCH models are simple and easy to handle
  + ARCH models take care of clustered errors
  + ARCH models take care of nonlinearities
  + ARCH models take care of changes in the econometrician’s ability to forecast
* The econometric challenge is to specify how the information is used to forecast the mean and variance of the variable of interest, conditional on the past information
* ARCH model assumes that the variance of tomorrow’s return is an equally weighted average of the squared residuals from the past period.
* The ARCH(1) model is specified as:
  + Yt = βt + ϵt ⋅ ut
  + where: ϵt = (α0 + α1ϵ2t−1)1/2

**Codes:**

library(AER)

data("Fatalities")

is.data.frame(Fatalities)

dim(Fatalities)

str(Fatalities)

head(Fatalities)

summary(Fatalities[, c(1, 2)])

# define the fatality rate

Fatalities$fatal\_rate <- Fatalities$fatal / Fatalities$pop \* 10000

# subset the data

Fatalities1982 <- subset(Fatalities, year == "1982")

Fatalities1983 <- subset(Fatalities, year == "1983")

Fatalities1988 <- subset(Fatalities, year == "1988")

Fatalities1987 <- subset(Fatalities, year == "1987")

plot(x = Fatalities1983$beertax,

y = Fatalities1983$fatal\_rate,

xlab = "Beer tax (in 1988 dollars)",

ylab = "Fatality rate (fatalities per 10000)",

main = "Traffic Fatality Rates and Beer Taxes in 1983",

ylim = c(0, 4.5),

pch = 20,

col = "steelblue")

abline(fatal1983\_mod, lwd = 1.5)

plot(x = Fatalities1987$beertax,

y = Fatalities1987$fatal\_rate,

xlab = "Beer tax (in 1988 dollars)",

ylab = "Fatality rate (fatalities per 10000)",

main = "Traffic Fatality Rates and Beer Taxes in 1987",

ylim = c(0, 4.5),

pch = 20,

col = "steelblue")

fatal1983\_mod <- lm(fatal\_rate ~ beertax, data = Fatalities1983)

coeftest(fatal1983\_mod, vcov. = vcovHC, type = "HC1")

fatal\_tefe\_lm\_mod <- lm(fatal\_rate ~ beertax + state + year - 1, data = Fatalities)

fatal\_tefe\_lm\_mod

coeftest(fatal\_tefe\_lm\_mod, vcov = vcovHC, type = "HC1")