Traffic Deaths and Beer Taxes: Panel Data Analysis and Instrumental Variables

Econ 440 - Introduction to Econometrics

Patrick Toche, ptoche@fullerton.edu

19 May 2022

Traffic Deaths and Beer Taxes

We use Christopher Ruhm's dataset for the period 1982-1988 to model the relationship between the beer tax and the traffic fatality rate, measured as the number of fatalities per $10,000$ inhabitants.

Ruhm, C. J. (1996). "Alcohol Policies and Highway Vehicle Fatalities." *Journal of Health Economics*, 15, 435–454.

Regression models may suffer from problems like omitted variables, measurement errors and simultaneous causality, causing the error term to be correlated with the regressor of interest, with the result that the least squares estimator is inconsistent. Adding omitted variables to the regression and using panel data entity and time fixed effects techniques can reduce estimation bias. However, if there is simultaneous causality (causality runs from X to Y and from Y to X), instrumental variables (IV) regression may be more suitable to obtain consistent estimates. We explore these issues in this notebook.

The State Traffic Fatality Data Set:

The data are for the contiguous 48 U.S. states (excluding Alaska and Hawaii), annually for 1982 through 1988. The traffic fatality rate is the number of traffic deaths in a given state in a given year per $10,000$ people living in that state in that year. Traffic fatality data were obtained from the *U.S. Department of Transportation Fatal Accident Reporting System.* The beer tax (the tax on a case of beer) was obtained from *Beer Institute's Brewers Almanac.* The beer tax is expressed in 1988 dollars. The drinking age variables are binary variables indicating whether the legal drinking age is 18, 19, or 20. The binary punishment variable describes the state's minimum sentencing requirements for an initial drunk driving conviction: $jail =$ ”*yes*” if the state requires jail time or community service and $jail =$ ”*no*” otherwise. Data on the total vehicle miles traveled annually by state were obtained from the *Department of Transportation.* Personal income data were obtained from the *U.S. Bureau of Economic Analysis*, and the unemployment rate was obtained from the *U.S. Bureau of Labor Statistics.* These data were graciously provided by Professor Christopher J. Ruhm of the Department of Economics at the University of North Carolina.

The dataset consists of 336 observations on 34 variables. The variable *state* is a factor variable with 48 levels. The variable *year* is a factor variable with 7 levels identifying the year when the observation was made. This gives $7 \times 48 = 336$ observations in total. Since all variables are observed for all entities and over all time periods, the panel is *balanced.*

Data preparation:

Convert binary variables to factors:

```
df$state <- factor(df$state)
df$year <- factor(df$year)
```

Table 1: **Description of Variables**

Name	Type	Description
state	factor	State ID (USPS Code)
year	factor	Year
spirits	numeric	Spirits Consumption
unemp	numeric	Unemployment Rate (Percent)
income	numeric	Per Capita Personal Income in 1987 dollars
emppop	numeric	Employment/Population Ratio
beertax	numeric	Tax on Case of Beer
baptist	numeric	Southern Baptist (Percent)
mormon	numeric	Mormon (Percentage)
drinkage	numeric	Minimum Legal Drinking Age
dry	numeric	Residing in Dry Counties (Percent)
youngdrivers	numeric	Drivers Aged 15-24 (Percent)
miles	numeric	Average Mile per Driver
breath	factor	Preliminary Breath Test Law
jail	factor	Mandatory Jail Sentence
service	factor	Mandatory Community Service
fatal	numeric	Number of Vehicle Fatalities
nfatal	numeric	Number of Night-Time Vehicle Fatalities
sfatal	numeric	Number of Single-Vehicle Fatalities
fatal1517	numeric	Number of Vehicle Fatalities Aged 15-17
nfatal1517	numeric	Number of Night-Time Vehicle Fatalities Aged 15-17
fatal1820	numeric	Number of Vehicle Fatalities Aged 18-20
nfatal1820	numeric	Number of Night-Time Vehicle Fatalities Aged 18-20
fatal2124	numeric	Number of Vehicle Fatalities Aged 21-24
nfatal2124	numeric	Number of Night-Time Vehicle Fatalities Aged 21-24
afatal	numeric	Number of Alcohol-Involved Vehicle Fatalities
pop	numeric	Population
pop1517	numeric	Population Aged 15-17
pop1820	numeric	Population Aged 18-20
pop2124	numeric	Population Aged 21-24
milestot	numeric	Total Vehicle Miles (Millions)
unempus	numeric	US Unemployment Rate
emppopus	numeric	US Employment/Population Ratio
gsp	numeric	Gross State Product (GSP) Rate of Change

```
df$breath <- factor(df$breath, levels=c("yes", "no"), labels=c(TRUE, FALSE))
df$jail <- factor(df$jail, levels=c("yes", "no"), labels=c(TRUE, FALSE))
df$service <- factor(df$service, levels=c("yes", "no"), labels=c(TRUE, FALSE))
```

Define the fatality rate per $10,000$:

```
df$fatality <- df$fatal / df$pop * 10000
```

Question 1.

Causality

Why would an increase in the beer tax be expected to reduce traffic fatalities? Detail the potential causality channels.

By raising the price of beer, beer taxes discourage drinking, thereby reducing drink-driving and fatalities.

Question 2.

Pooled Linear Regression With No Controls

Estimate a simple linear regression model of the effect of the beer tax on the fatality rate, with no controls. Report the estimated coefficient and standard error. Is the coefficient significant at the 0.05 significance level? Does the coefficient have the expected sign? Comment.

```
lm(fatality ~ beertax, data=df) %>% summary()
```

```
## 
## Call:
## lm(formula = fatality ~ beertax, data = df)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1.0906 -0.3777 -0.0944  0.2855  2.2764 
## 
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)    
## (Intercept)   1.8533     0.0436   42.54 < 0.0000000000000002 ***
## beertax       0.3646     0.0622    5.86          0.000000011 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.544 on 334 degrees of freedom
## Multiple R-squared:  0.0934, Adjusted R-squared:  0.0906 
## F-statistic: 34.4 on 1 and 334 DF,  p-value: 0.0000000108
```

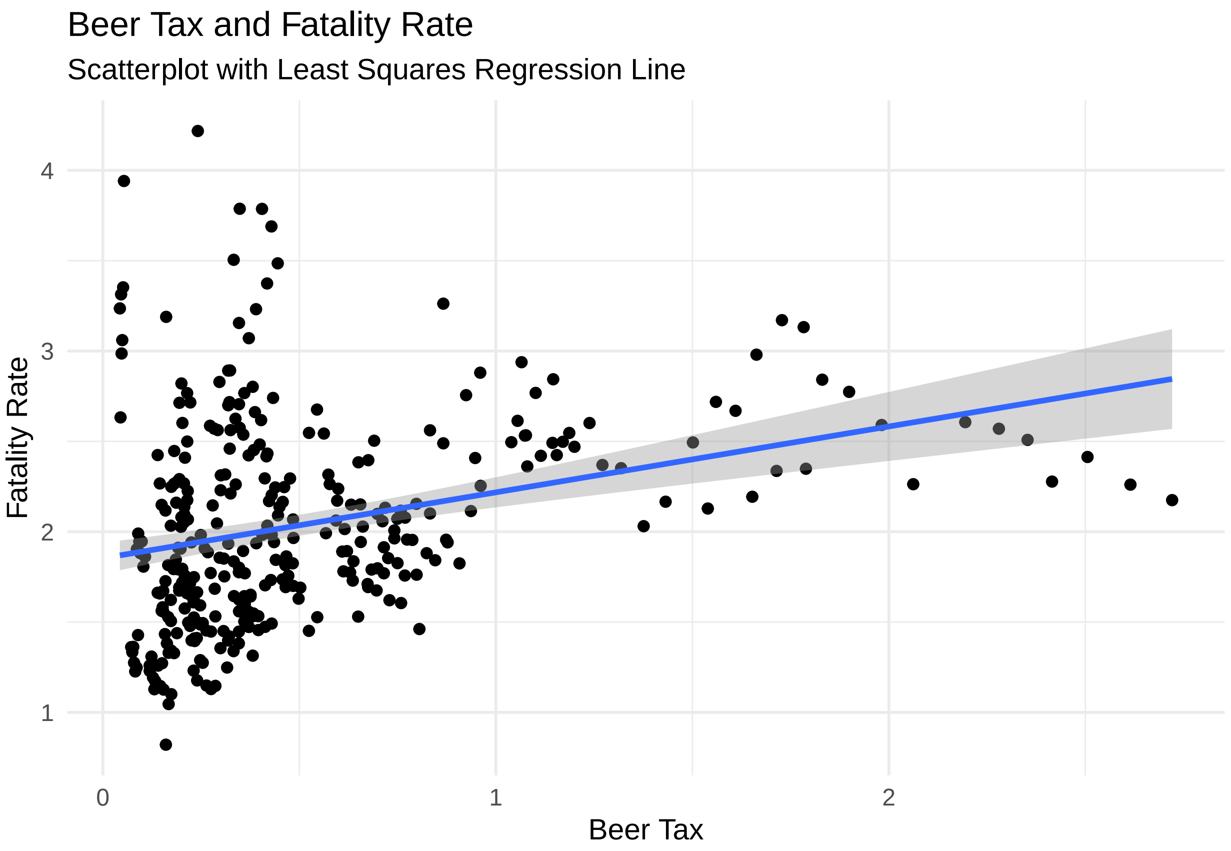
The estimated coefficient is 0.3646 and standard error 0.0622. The estimate is significant at the 0.05 significance level. But the positive sign is not what was expected.

Question 3.

Scatterplot with Pooled Regression Lines

Produce a scatterplot of the fatality rate and the beer tax.

```
ggplot(df, aes(x=beertax, y=fatality)) +
  geom_point() +
  geom_smooth(method="lm") +
  theme_minimal() +
  labs(title="Beer Tax and Fatality Rate",
       x="Beer Tax",
       y="Fatality Rate",
       subtitle="Scatterplot with Least Squares Regression Line")
```



Question 4.

Linear Regression Model For Each Year

To isolate potential shifts in the population regression parameters, estimate a simple linear regression model for each year in the dataset. What is the range of the estimated coefficients? Do the coefficients have the expected sign?

One approach is to subset the data to the years of interest and estimate a linear model for each subset:

```
tidy(m82)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    2.01      0.139    14.5   1.01e-18
## 2 beertax        0.148     0.188     0.788 4.35e- 1
```

```
tidy(m83)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    1.85      0.124     14.9  2.86e-19
## 2 beertax        0.299     0.168      1.78 8.22e- 2
```

```
tidy(m84)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    1.81      0.110     16.5  6.53e-21
## 2 beertax        0.400     0.152      2.64 1.14e- 2
```

```
tidy(m85)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    1.77      0.109     16.3  9.67e-21
## 2 beertax        0.392     0.155      2.53 1.48e- 2
```

```
tidy(m86)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    1.82      0.111     16.4  7.66e-21
## 2 beertax        0.480     0.161      2.98 4.64e- 3
```

```
tidy(m87)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    1.82      0.113     16.1  1.63e-20
## 2 beertax        0.483     0.170      2.85 6.58e- 3
```

```
tidy(m88)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    1.86      0.106     17.5  5.28e-22
## 2 beertax        0.439     0.164      2.67 1.05e- 2
```

More elegant and quite straightforward is to leverage the power of the *tidy*() function from the “broom“ package and the $group_by()$ function of the “dplyr“ package:

```
df %>%
  group_by(year) %>%
  do(tidy(lm(fatality ~ beertax, .)))
```

```
## # A tibble: 14 x 6
## # Groups:   year [7]
##    year  term        estimate std.error statistic  p.value
##    <fct> <chr>          <dbl>     <dbl>     <dbl>    <dbl>
##  1 1982  (Intercept)    2.01      0.139    14.5   1.01e-18
##  2 1982  beertax        0.148     0.188     0.788 4.35e- 1
##  3 1983  (Intercept)    1.85      0.124    14.9   2.86e-19
##  4 1983  beertax        0.299     0.168     1.78  8.22e- 2
##  5 1984  (Intercept)    1.81      0.110    16.5   6.53e-21
##  6 1984  beertax        0.400     0.152     2.64  1.14e- 2
##  7 1985  (Intercept)    1.77      0.109    16.3   9.67e-21
##  8 1985  beertax        0.392     0.155     2.53  1.48e- 2
##  9 1986  (Intercept)    1.82      0.111    16.4   7.66e-21
## 10 1986  beertax        0.480     0.161     2.98  4.64e- 3
## 11 1987  (Intercept)    1.82      0.113    16.1   1.63e-20
## 12 1987  beertax        0.483     0.170     2.85  6.58e- 3
## 13 1988  (Intercept)    1.86      0.106    17.5   5.28e-22
## 14 1988  beertax        0.439     0.164     2.67  1.05e- 2
```

There is a positive correlation between beer tax and fatality rate. This is likely because of an omitted variable bias. Factors such as the age of the cars, the state of roads, road safety regulation, policing, laws, the local culture of drinking, traffic density, social problems, the level of local taxes, would be expected to affect fatality rate. Omitted variables cause least squares estimates to be biased. To the extent that unobserved variables are constant across time, panel data techniques can be used to reduce the bias.

Question 5.

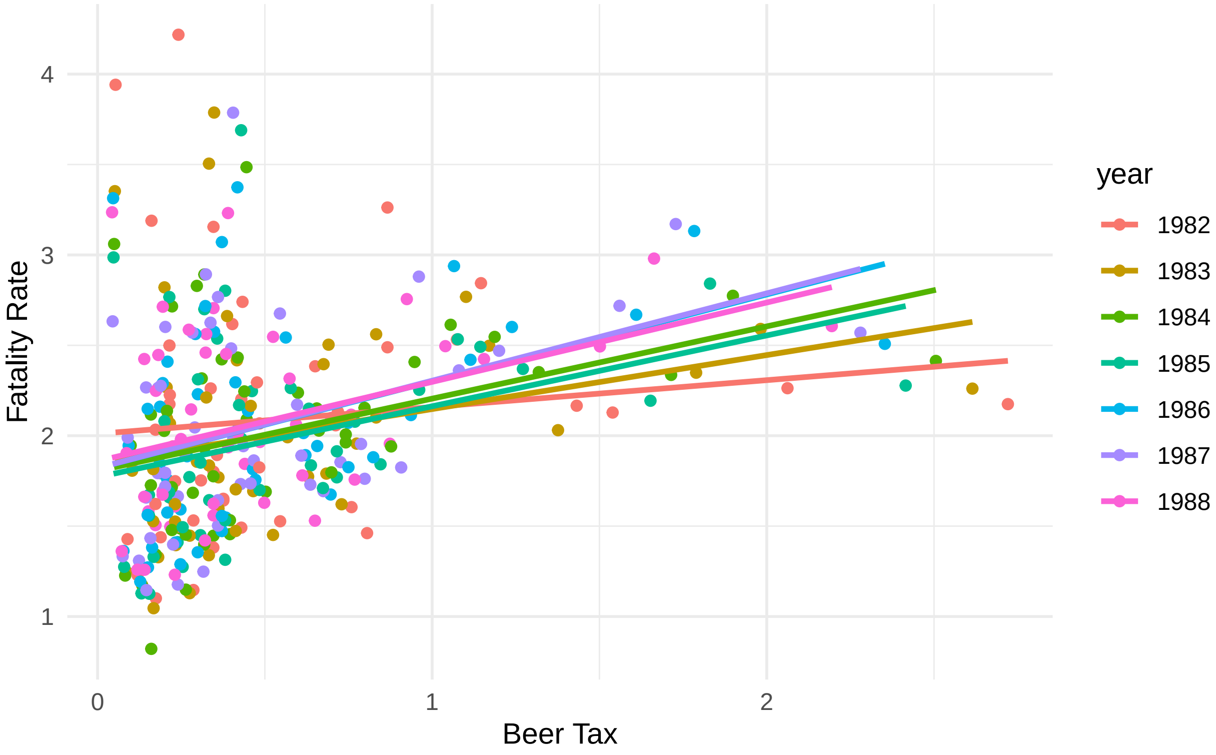
Scatterplot with Grouped Regression Lines

Produce a scatterplot of the fatality rate and the beer tax with a regression line for each year in the sample. Do the regression lines have the expected slope?

```
ggplot(df, aes(x=beertax, y=fatality, group=year, color=year)) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE) +
  theme_minimal() +
  labs(title="Beer Tax and Fatality Rate",
       x="Beer Tax",
       y="Fatality Rate",
       subtitle="Scatterplot with Least Squares Regression Line")
```

Beer Tax and Fatality Rate

Scatterplot with Least Squares Regression Line



Question 6.

Threats to Internal Validity

The five potential threats to the internal validity of a regression study are: omitted variables, misspecification of the functional form, imprecise measurement of the independent variables, sample selection, and simultaneous causality. Discuss how each of these potential issues may or may not be relevant to this study. Are there important omitted variables that affect traffic fatalities and that may be correlated with the other variables included in the regression? How would these issues affect least squares estimates?

- The most obvious candidates are the safety of roads, weather, and so forth. These variables are essentially constant over the sample period, so their effect is captured by the state fixed effects.

- Since most of the variables are binary variables, the largest functional form choice involves the Beer Tax variable.

- A linear specification is used in the text. To check the reliability of the linear specification, it would be useful to consider a log specification and a quadratic specification.

- Measurement error does not appear to be a problem, because variables like traffic fatalities, taxes, driving age, laws, and unemployment are all accurately measured.

- Sample selection does not appear to be a problem, because data was collected from all contiguous states in the continental United States.

- Simultaneous causality is a potential problem. States with high fatality rates could set higher beer taxes in an effort to reduce alcohol consumption and serious traffic accidents.

Question 7.

Pooled Regression Model With Controls

State laws and economic conditions are likely correlated with the prevalence of drunk-driving. Vehicle use depends on economic conditions: In recessions and/or when gas prices are high, vehicle owners will tend to drive less. States have different laws that target driving under the influence. Omitting these laws could produce omitted variable bias, even in regressions with state and time fixed effects. We therefore include control variables for driving laws and economic conditions.

Estimate a least squares regression with pooled data, using the following control variables: “spirits consumption”, “gross state product”, “minimum drinking age”, and binary variables for “preliminary breath test law”, “mandatory jail sentence”, “mandatory community service”, “dry county” (a dry county is a county that prohibits the sale of any kind of alcoholic beverages).

The most straightforward approach is to use the built-in *lm*() function.

```
lm(fatality ~ beertax + spirits + gsp + drinkage + breath + jail + service + dry, data=df) %>% tidy()
```

```
## # A tibble: 9 x 5
##   term         estimate std.error statistic      p.value
##   <chr>           <dbl>     <dbl>     <dbl>        <dbl>
## 1 (Intercept)   2.13      0.677       3.15  0.00178
## 2 beertax       0.355     0.0604      5.87  0.0000000106
## 3 spirits       0.0827    0.0447      1.85  0.0653
## 4 gsp          -2.81      0.682      -4.11  0.0000495
## 5 drinkage     -0.0111    0.0320     -0.346 0.729
## 6 breathFALSE   0.183     0.0603      3.04  0.00259
## 7 jailFALSE    -0.297     0.0792     -3.75  0.000213
## 8 serviceFALSE -0.0611    0.0867     -0.705 0.481
## 9 dry           0.00955   0.00312     3.06  0.00238
```

An alternative is to set the *model* = ”*pooling*” argument in the *plm* function from the “plm“ library:

```
plm(fatality ~ beertax + spirits + gsp + drinkage + breath + jail + service + dry,
    data=df, model="pooling") -> plm.pooled
tidy(plm.pooled)
```

```
## # A tibble: 9 x 5
##   term         estimate std.error statistic      p.value
##   <chr>           <dbl>     <dbl>     <dbl>        <dbl>
## 1 (Intercept)   2.13      0.677       3.15  0.00178
## 2 beertax       0.355     0.0604      5.87  0.0000000106
## 3 spirits       0.0827    0.0447      1.85  0.0653
## 4 gsp          -2.81      0.682      -4.11  0.0000495
## 5 drinkage     -0.0111    0.0320     -0.346 0.729
## 6 breathFALSE   0.183     0.0603      3.04  0.00259
## 7 jailFALSE    -0.297     0.0792     -3.75  0.000213
## 8 serviceFALSE -0.0611    0.0867     -0.705 0.481
## 9 dry           0.00955   0.00312     3.06  0.00238
```

Question 8.

First Difference Regression Model (difference between 1988 and 1982) with No Controls

If unobserved state-level characteristics are related to how the local beer tax is set, the least squares estimate will be biased and inconsistent. One way around this problem is to estimate a model in first differences, that is to consider the difference between the first and last years in the sample. If there are characteristics of a state that do not change over time, say between 1982 and 1988, any fixed effects cancel out after taking the difference. And we do not even need to know what these effects were, since they're gone!

Side-by-side comparison of 1982 and 1988 cross-section regressions:

```
stargazer(m82, m88,
          se=list(sqrt(diag(vcovHC(m82, type="HC3"))),
                  sqrt(diag(vcovHC(m88, type="HC3")))),
          title="Cross-Section Regressions for 1982 and 1988",
          type="text",
          column.labels=c("1982", "1988"),
          df=FALSE,
          digits=4)
```

```
##
## Cross-Section Regressions for 1982 and 1988
## ================================================
##                         Dependent variable:
##                     ----------------------------
##                               fatality
##                          1982          1988
##                          (1)            (2)
## ------------------------------------------------
## beertax                 0.1485       0.4388***
##                        (0.1450)      (0.1422)
##
## Constant               2.0104***     1.8591***
##                        (0.1528)      (0.1179)
##
## ------------------------------------------------
## Observations              48            48
## R2                      0.0133        0.1340
## Adjusted R2            -0.0081        0.1152
## Residual Std. Error     0.6705        0.4903
## F Statistic             0.6212       7.1180**
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Take a first difference across two time periods and estimate a regression using the differenced data:

```
fatality.diff <- df88$fatality - df82$fatality
beertax.diff <- df88$beertax - df82$beertax
lm.diff <- lm(fatality.diff ~ beertax.diff)
tidy(lm.diff)
```

```
## # A tibble: 2 x 5
##   term         estimate std.error statistic p.value
##   <chr>           <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)   -0.0720    0.0606     -1.19   0.241
## 2 beertax.diff  -1.04      0.417      -2.49   0.0162
```

Estimate robust standard errors:

```
coeftest(lm.diff, vcov=vcovHC, type="HC3")
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.0720     0.0679   -1.06    0.294
## beertax.diff  -1.0410     0.4083   -2.55    0.014 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Even without controls, the sign of the estimated coefficient is now negative, as our initial conjecture suggests.

Question 9.

First Difference Regression Model (difference between 1988 and 1982) With Controls

The ‘`plm` library provides the convenient *model* = ”*fd*” argument to the *plm*() function to compute a first-difference (fd) regression. Estimate the regression in first-differences using the set of control variables identified earlier.

Check that the *plm*() function replicates the first-difference model estimated above:

```
plm(fatality ~ beertax, data=bind_rows(df82,df88),
    index=c("state", "year"),
    model="fd") %>% tidy()
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)  -0.0720    0.0606     -1.19  0.241
## 2 beertax      -1.04      0.417      -2.49  0.0162
```

Adding control variables to the first-difference regression reduces the size of the coefficient on *beertax*:

```
plm.diff <- plm(fatality ~ beertax + spirits + gsp + drinkage + breath + jail + service + dry,
                data=bind_rows(df82,df88),
                index=c("state", "year"),
                model="fd")
tidy(plm.diff)
```

```
## # A tibble: 8 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)  -0.137     0.137     -0.996 0.325
## 2 beertax      -0.873     0.310     -2.81  0.00766
## 3 spirits       1.19      0.250      4.75  0.0000273
## 4 gsp           6.52      1.51       4.31  0.000107
## 5 drinkage      0.0778    0.0419     1.86  0.0707
## 6 breathFALSE  -0.122     0.104     -1.18  0.247
## 7 jailFALSE    -0.0307    0.146     -0.211 0.834
## 8 dry           0.0242    0.0304     0.798 0.430
```

Question 10.

Identifying Assumptions of First Difference Model

We review the key identifying assumptions for the first-difference estimator.

1) By differencing out all the time-independent variation in the independent variables, we have also reduced the amount of variation that can be used for identification, which leads to greater standard errors. To reduce imprecision in the estimates, the best approach is to increase the sample size.

2) Bias that arises from measurement error is irreducible and could be increased by the first-difference estimator if stochastic noise varies with time. Measurement error cause bias and leads to greater standard errors.

3) Omitted variable bias is not eliminated by the first-difference estimator.

4) If differenced-errors are autocorrelated, the first-difference estimator can be imprecise. Autocorrelation can arise from the process of first-differencing.

If the errors are autocorrelated, the conventional estimates of the standard errors are incorrect. Report clustered standard errors for the first-difference model.

```
coeftest(plm.diff, vcov=vcovHC, type="HC3")
```

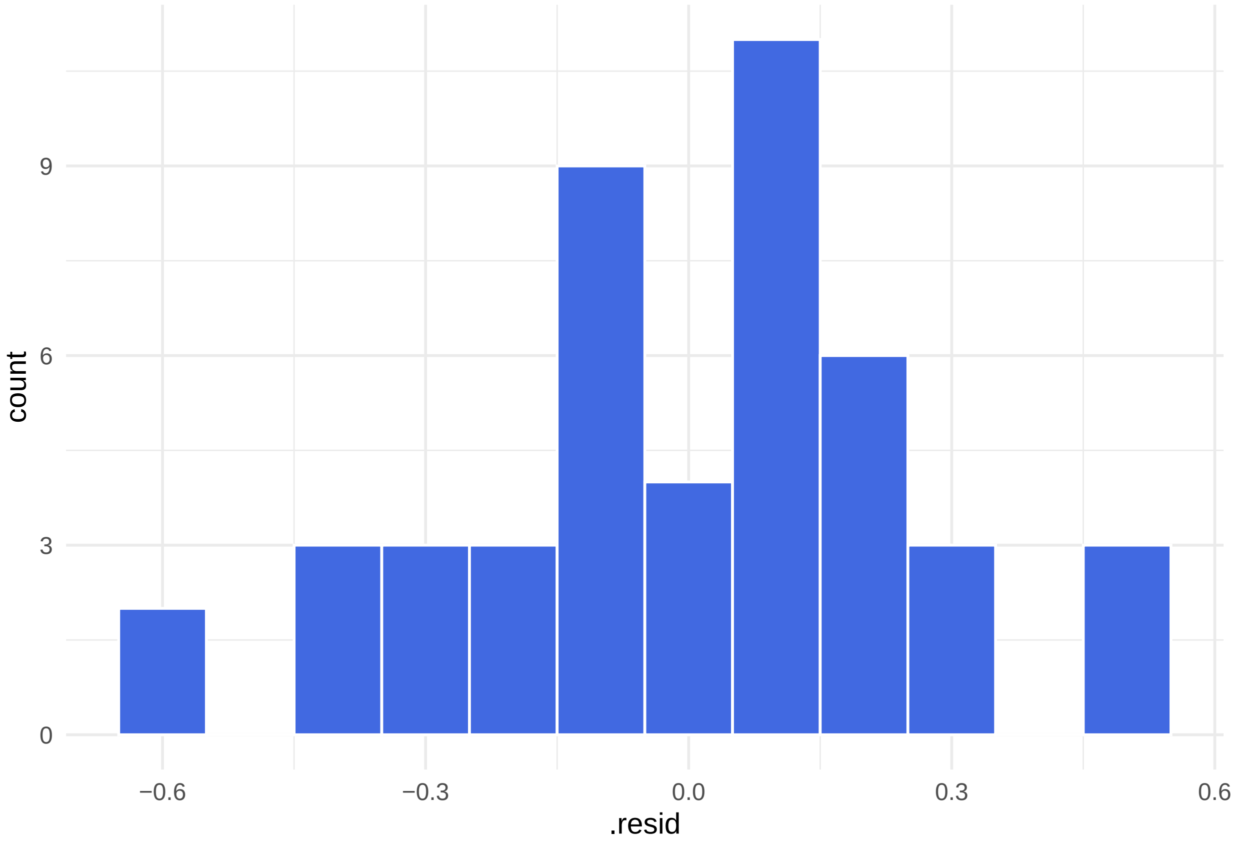
```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.1367     0.1588   -0.86  0.39481
## beertax      -0.8729     0.3409   -2.56  0.01442 *
## spirits       1.1895     0.3009    3.95  0.00032 ***
## gsp           6.5221     1.6804    3.88  0.00039 ***
## drinkage      0.0778     0.0445    1.75  0.08830 .
## breathFALSE  -0.1217     0.1229   -0.99  0.32847
## jailFALSE    -0.0307     0.1616   -0.19  0.85034
## dry           0.0242     0.0448    0.54  0.59213
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Store the residuals and fitted values:

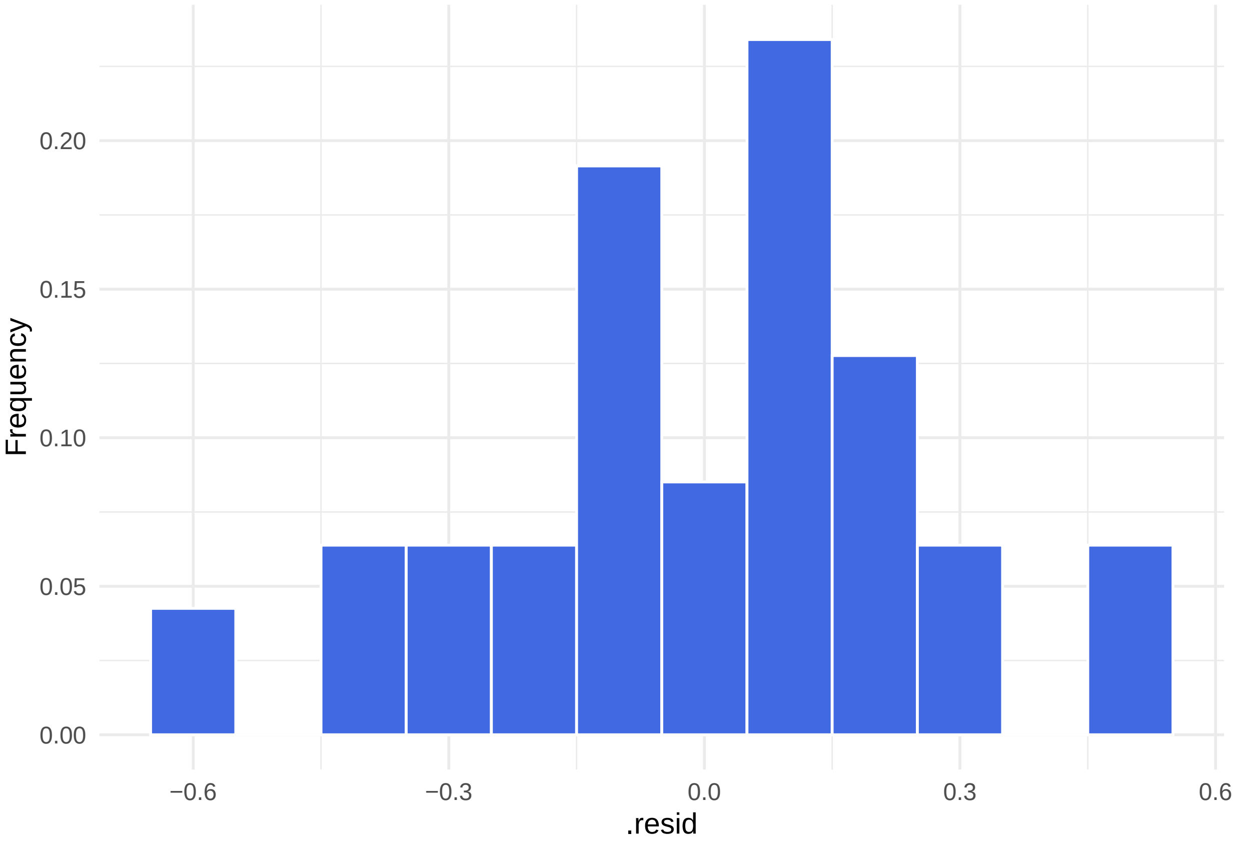
```
data.frame(".rownames" = row.names(plm.diff$model), plm.diff$model) %>%
  left_join(data.frame(".rownames" = names(resid(plm.diff)),
                       ".fitted" = fitted(plm.diff),
                       ".resid" = resid(plm.diff)
                       )) %>% na.omit() -> df.diff
```

```
## Joining, by = ".rownames"
```

```
df.diff %>%
  ggplot(.,aes(x=.resid)) +
  geom_histogram(fill="royalblue", col="white", binwidth=0.10) +
  theme_minimal()
```

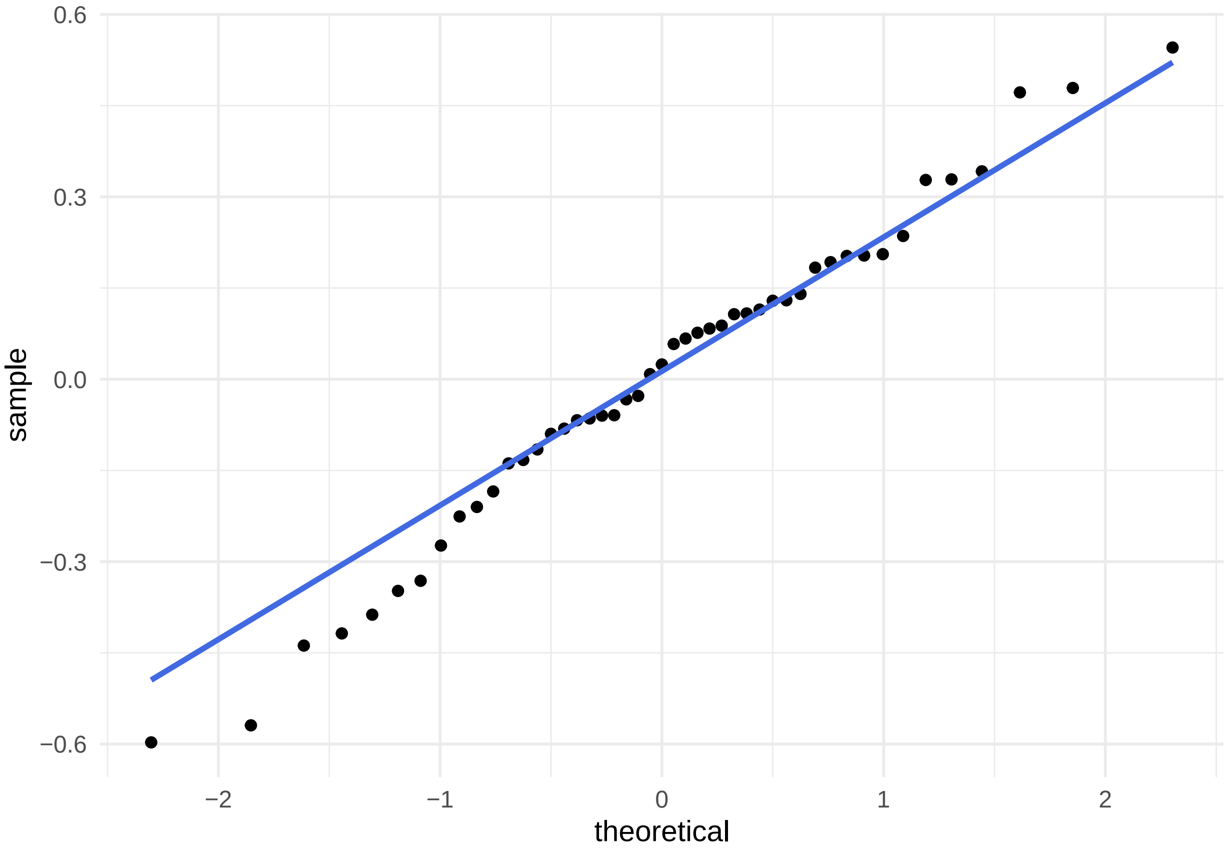


```
df.diff %>%
  ggplot(.,aes(x = .resid)) +
  geom_histogram(fill="royalblue", col="white", binwidth=0.10,
                 aes(y=..count../sum(..count..))) +
  labs(y="Frequency") +
  theme_minimal()
```



Examine the residuals with a qq plot:

```
# qqnorm(resid(plm.diff))  # base R qq-plot
ggplot(aes(sample=.resid), data=df.diff) +
  geom_qq() +
  geom_qq_line(col="royalblue", size=1) +
  labs(x="theoretical", y="sample") +
  theme_minimal()
```



```
stargazer(plm.pooled, plm.diff, plm.diff,
          se=list(NULL, NULL, sqrt(diag(vcovHC(plm.diff, type="HC3")))),
          title="Panel Regression with Fixed Effecs", type="text",
          column.labels=c("Pooled OLS", "FE Conventional Standard Errors", "FE Clustered Standard Error
          df=FALSE,
          digits=4)
```

```
##
## Panel Regression with Fixed Effecs
## ====================================================================================
##                                         Dependent variable:
##              -----------------------------------------------------------------------
##                                              fatality
##              Pooled OLS FE Conventional Standard Errors FE Clustered Standard Errors
##                 (1)                   (2)                           (3)
## ------------------------------------------------------------------------------------
## beertax      0.3546***            -0.8729***                     -0.8729**
##               (0.0604)             (0.3104)                       (0.3409)
##
## spirits       0.0827*              1.1895***                     1.1895***
##               (0.0447)             (0.2503)                       (0.3009)
##
## gsp          -2.8055***            6.5221***                     6.5221***
##               (0.6821)             (1.5133)                       (1.6804)
##
## drinkage      -0.0111               0.0778*                       0.0778*
##               (0.0320)             (0.0419)                       (0.0445)
```

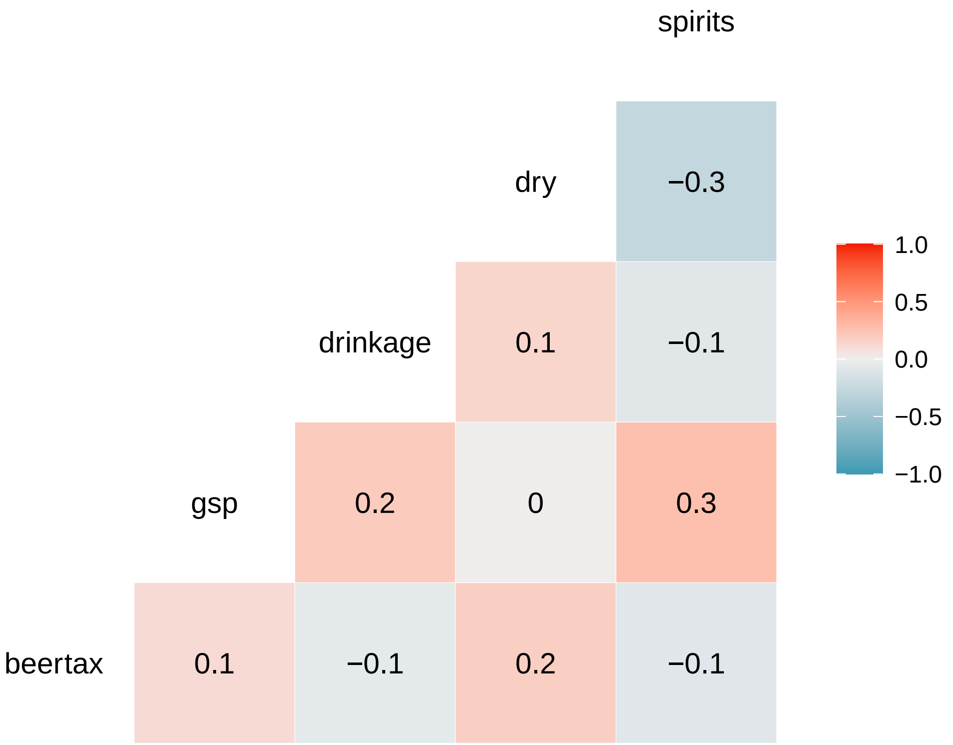
```
##
## breathFALSE  0.1830***              -0.1217                       -0.1217
##               (0.0603)             (0.1035)                       (0.1229)
##
## jailFALSE    -0.2965***             -0.0307                       -0.0307
##               (0.0792)             (0.1456)                       (0.1616)
##
## serviceFALSE  -0.0611
##               (0.0867)
##
## dry          0.0095***              0.0242                         0.0242
##               (0.0031)             (0.0304)                       (0.0448)
##
## Constant     2.1348***              -0.1367                       -0.1367
##               (0.6774)             (0.1372)                       (0.1588)
##
## ------------------------------------------------------------------------------------
## Observations    335                    47                             47
## R2               0.2616               0.6094                         0.6094
## Adjusted R2      0.2435               0.5392                         0.5392
## F Statistic  14.4390***            8.6910***                     8.6910***
## ====================================================================================
## Note:                                                    *p<0.1; **p<0.05; ***p<0.01
```

Question 11

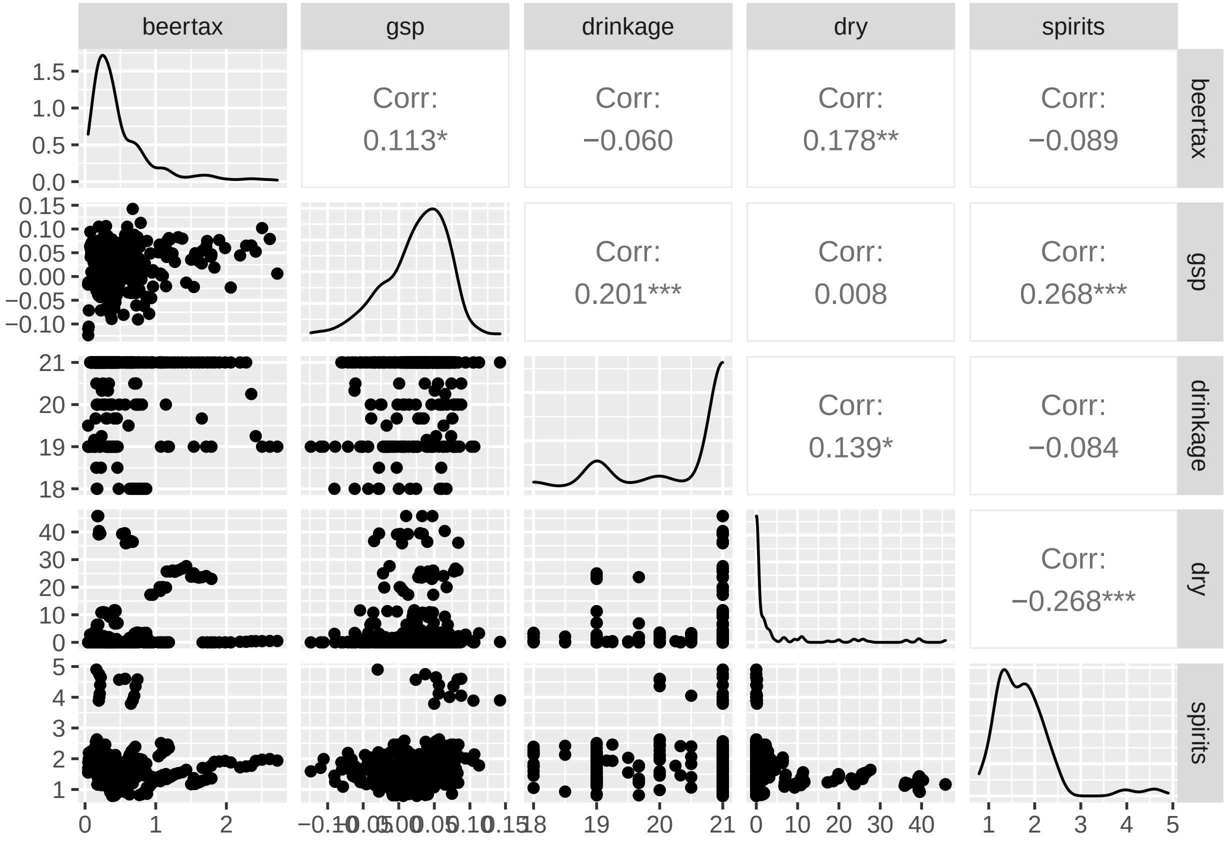
Multicollinearity in the First Difference Model

There are many factors that influence traffic safety, and if they change over time and are correlated with the real beer tax, then their omission will produce omitted variable bias. If the fixed effect is correlated with the independent variables, the estimated coefficient will be biased.

```
df.corr <- df[, c("beertax", "gsp", "drinkage", "dry", "spirits")]
ggcorr(df.corr, method=c("everything", "pearson"),  label=TRUE)
```



```
ggpairs(df.corr)
```



Question 12

Fixed Effect Regression

The fixed effects model is one of the simplest and most robust specifications in panel data econometrics and often used as a benchmark against which more sophisticated techniques are compared. Estimate the fixed effects regression model:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + \beta_3 S_t + u_{it}, \quad i = 1, 2, \ldots, 48; \quad t = 1982, \ldots, 1988$$

where:

Y_{it} is the fatality rate (per $10,000$ individuals) for state i at time t; X_{it} are the regressors; Z_i are state fixed effects; S_t are time fixed effect; uit is the error term.

Assume that u_{it} are zero-mean normal random variables; large outliers are unlikely (X_{it}, u_{it}) have finite 4th moments); covariants are independent across states. Several state characteristics associated with traffic fatalities are used as control variables. Consider the following: "spirits consumption", "gross state product", "minimum drinking age", and binary variables for "preliminary breath test law", "mandatory jail sentence", "mandatory community service", "dry county" (as before); and "per capita personal income", and the "state unemployment rate".

Which coefficients are significant at the 0.05 significance level? Which are not? Is there evidence for entity fixed effects? Is there evidence for time fixed effects? How does the estimated coefficient on *beertax* change by including time fixed effects?

Pooled Panel Regression No Fixed Effects:

```
formula1 <- "fatality ~ beertax + spirits + gsp + drinkage + breath + jail + service + dry + income + u
formula2 <- "fatality ~ beertax + gsp + income + unemp"
plm(formula1,
    data = df,
    index = c("state", "year"),
    model = "pooling") -> plm.pooled
coeftest(plm.pooled, vcov=vcovHC, type="HC3")
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.8351383  1.1627586    3.30  0.00108 **
## beertax       0.0791470  0.0957182    0.83  0.40892
## spirits       0.2142804  0.1011968    2.12  0.03498 *
## gsp          -1.1730522  1.4206169   -0.83  0.40956
## drinkage      0.0184340  0.0610396    0.30  0.76284
## breathFALSE   0.2723137  0.1250740    2.18  0.03019 *
## jailFALSE    -0.1349839  0.2086963   -0.65  0.51822
## serviceFALSE -0.1354602  0.1864924   -0.73  0.46814
## dry           0.0034105  0.0052148    0.65  0.51358
## income       -0.0001568  0.0000414   -3.79  0.00018 ***
## unemp        -0.0457950  0.0248187   -1.85  0.06592 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Entity Fixed Effects but NO Time Fixed Effects:

```
plm(formula1,
    data = df,
    index = c("state", "year"),
    model = "within",
    effect = "individual"
    ) -> plm.fe
coeftest(plm.fe, vcov=vcovHC, type="HC3")
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error t value     Pr(>|t|)    
## beertax      -0.4713276  0.2526205   -1.87       0.0631 .  
## spirits       0.8507160  0.1421251    5.99 0.0000000067 ***
## gsp          -0.5054940  0.2345748   -2.15       0.0320 *  
## drinkage      0.0215798  0.0227410    0.95       0.3435    
## breathFALSE  -0.0181567  0.0493288   -0.37       0.7131    
## jailFALSE    -0.0121961  0.0109525   -1.11       0.2664    
## serviceFALSE  0.0082110  0.1488115    0.06       0.9560    
## dry           0.0257081  0.0127651    2.01       0.0450 *  
## income        0.0000981  0.0000338    2.90       0.0040 ** 
## unemp        -0.0352517  0.0107577   -3.28       0.0012 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Entity Fixed Effects and Time Fixed Effects

```
plm(formula1,
    data = df,
    index = c("state", "year"),
    model = "within",
    effect = "twoways"
    ) -> plm.fe.te
coeftest(plm.fe.te, vcov=vcovHC, type="HC3")
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error t value     Pr(>|t|)    
## beertax      -0.4631708  0.2648379   -1.75       0.0814 .  
## spirits       0.8314525  0.1235116    6.73 0.0000000001 ***
## gsp           0.3567483  0.3543051    1.01       0.3149    
## drinkage      0.0142885  0.0217297    0.66       0.5114    
## breathFALSE  -0.0306647  0.0399871   -0.77       0.4438    
## jailFALSE    -0.0496327  0.0163228   -3.04       0.0026 ** 
## serviceFALSE  0.0147757  0.1445126    0.10       0.9186    
## dry           0.0190566  0.0102359    1.86       0.0637 .  
## income        0.0000839  0.0000319    2.63       0.0091 ** 
## unemp        -0.0514896  0.0130541   -3.94       0.0001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test if fixed effects, particularly fixed time effects, are significant:

```
pFtest(fatality ~ beertax + spirits + gsp + drinkage + breath + jail + service + dry + income + unemp,
```

```
##
##  F test for twoways effects
```

```
##
## data:  fatality ~ beertax + spirits + gsp + drinkage + breath + jail +  ...
## F = 51.6, df1 = 53, df2 = 271, p-value <0.0000000000000002
## alternative hypothesis: significant effects
```

```
pFtest(fatality ~ beertax + spirits + gsp + drinkage + breath + jail + service + dry + income + unemp,
```

```
##
##  F test for individual effects
##
## data:  fatality ~ beertax + spirits + gsp + drinkage + breath + jail +  ...
## F = 49.9, df1 = 47, df2 = 277, p-value <0.0000000000000002
## alternative hypothesis: significant effects
```

```
pFtest(fatality ~ beertax + spirits + gsp + drinkage + breath + jail + service + dry + income + unemp,
```

```
##
##  F test for time effects
##
## data:  fatality ~ beertax + spirits + gsp + drinkage + breath + jail +  ...
## F = 4.69, df1 = 6, df2 = 318, p-value = 0.00014
## alternative hypothesis: significant effects
```

Question 13

Explain the benefits of fixed effect models over the first-difference model.

The two-way fixed effect model eliminates bias from unobservables that change over time but are constant over entities and controls for factors that differ across entities but are constant over time.

Question 14

Fixed Effect Regression with Nonlinear Covariates

Estimate the previous regression replacing the level of "per capita personal income" with its logarithm. Explore other non-linear functions. Is there any evidence of non-linear effects in the other regressors, e.g. *beertax*? Comment.

Question 15

Several variables in the dataset capture policies intended to affect alcohol consumption. Because these laws can only affect motor vehicle fatalities via alcohol consumption (beer tax, minimum legal drinking age, dry county laws, preliminary breath test law, mandatory jail sentence, mandatory community service) they can used as instrumental variables (IVs) to estimate the effect of alcohol consumption (*spirits*). Estimate the effect of alcohol consumption on vehicle fatalities using alcohol laws as instrumental variables.