Regression with Panel Data: Traffic Deaths and Alcohol Taxes

Econ 440 - Introduction to Econometrics

Patrick Toche, ptoche@fullerton.edu

13 May 2022

Dataset:

```
#library(AER)
data(Fatalities)
df <- Fatalities
rm(Fatalities)
```

Data structure:

```
str(df)
```

```
## 'data.frame':    336 obs. of  34 variables:
##  $ state       : Factor w/ 48 levels "al","az","ar",..: 1 1 1 1 1 1 1 2 2 2 ...
##  $ year        : Factor w/ 7 levels "1982","1983",..: 1 2 3 4 5 6 7 1 2 3 ...
##  $ spirits     : num  1.37 1.36 1.32 1.28 1.23 ...
##  $ unemp       : num  14.4 13.7 11.1 8.9 9.8 ...
##  $ income      : num  10544 10733 11109 11333 11662 ...
##  $ emppop      : num  50.7 52.1 54.2 55.3 56.5 ...
##  $ beertax     : num  1.54 1.79 1.71 1.65 1.61 ...
##  $ baptist     : num  30.4 30.3 30.3 30.3 30.3 ...
##  $ mormon      : num  0.328 0.343 0.359 0.376 0.393 ...
##  $ drinkage    : num  19 19 19 19.7 21 ...
##  $ dry         : num  25 23 24 23.6 23.5 ...
##  $ youngdrivers: num  0.212 0.211 0.211 0.211 0.213 ...
##  $ miles       : num  7234 7836 8263 8727 8953 ...
##  $ breath      : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ jail        : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 2 2 2 ...
##  $ service     : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 2 2 2 ...
##  $ fatal       : int  839 930 932 882 1081 1110 1023 724 675 869 ...
##  $ nfatal      : int  146 154 165 146 172 181 139 131 112 149 ...
##  $ sfatal      : int  99 98 94 98 119 114 89 76 60 81 ...
##  $ fatal1517   : int  53 71 49 66 82 94 66 40 40 51 ...
##  $ nfatal1517  : int  9 8 7 9 10 11 8 7 7 8 ...
##  $ fatal1820   : int  99 108 103 100 120 127 105 81 83 118 ...
##  $ nfatal1820  : int  34 26 25 23 23 31 24 16 19 34 ...
##  $ fatal2124   : int  120 124 118 114 119 138 123 96 80 123 ...
##  $ nfatal2124  : int  32 35 34 45 29 30 25 36 17 33 ...
##  $ afatal      : num  309 342 305 277 361 ...
##  $ pop         : num  3942002 3960008 3988992 4021008 4049994 ...
##  $ pop1517     : num  209000 202000 197000 195000 204000 ...
##  $ pop1820     : num  221553 219125 216724 214349 212000 ...
```

```
##  $ pop2124     : num  290000 290000 288000 284000 263000 ...
##  $ milestot    : num  28516 31032 32961 35091 36259 ...
##  $ unempus     : num  9.7 9.6 7.5 7.2 7 ...
##  $ emppopus    : num  57.8 57.9 59.5 60.1 60.7 ...
##  $ gsp         : num  -0.0221 0.0466 0.0628 0.0275 0.0321 ...
```

```r
# check that state and year are factors
class(df$state)
```

```
## [1] "factor"
```

```r
class(df$year)
```

```
## [1] "factor"
```

Data slice:

```r
head(df)
```

```
##   state year spirits unemp income emppop beertax baptist  mormon drinkage
## 1    al 1982    1.37  14.4  10544 50.692  1.5394  30.356 0.32829    19.00
## 2    al 1983    1.36  13.7  10733 52.147  1.7890  30.334 0.34341    19.00
## 3    al 1984    1.32  11.1  11109 54.168  1.7143  30.312 0.35924    19.00
## 4    al 1985    1.28   8.9  11333 55.271  1.6525  30.289 0.37579    19.67
## 5    al 1986    1.23   9.8  11662 56.514  1.6099  30.267 0.39311    21.00
## 6    al 1987    1.18   7.8  11944 57.510  1.5600  30.245 0.41123    21.00
##      dry youngdrivers  miles breath jail service fatal nfatal sfatal fatal1517
## 1 25.006      0.21157 7233.9     no   no      no   839    146     99        53
## 2 22.994      0.21077 7836.3     no   no      no   930    154     98        71
## 3 24.043      0.21148 8263.0     no   no      no   932    165     94        49
## 4 23.634      0.21114 8726.9     no   no      no   882    146     98        66
## 5 23.465      0.21340 8952.9     no   no      no  1081    172    119        82
## 6 23.792      0.21553 9166.3     no   no      no  1110    181    114        94
##   nfatal1517 fatal1820 nfatal1820 fatal2124 nfatal2124 afatal     pop pop1517
## 1          9        99         34       120         32 309.44 3942002  209000
## 2          8       108         26       124         35 341.83 3960008  202000
## 3          7       103         25       118         34 304.87 3988992  197000
## 4          9       100         23       114         45 276.74 4021008  195000
## 5         10       120         23       119         29 360.72 4049994  204000
## 6         11       127         31       138         30 368.42 4082999  205000
##   pop1820 pop2124 milestot unempus emppopus       gsp
## 1  221553  290000    28516     9.7     57.8 -0.022125
## 2  219125  290000    31032     9.6     57.9  0.046558
## 3  216724  288000    32961     7.5     59.5  0.062798
## 4  214349  284000    35091     7.2     60.1  0.027490
## 5  212000  263000    36259     7.0     60.7  0.032143
## 6  208998  259000    37426     6.2     61.5  0.048976
```

Data summary for *state* and *year*:

```r
summary(df[, c(1, 2)])
```

```
##      state        year   
##  al     :  7   1982:48  
##  az     :  7   1983:48  
##  ar     :  7   1984:48  
```

```
##  ca      :  7   1985:48
##  co      :  7   1986:48
##  ct      :  7   1987:48
##  (Other):294   1988:48
```

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 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*.

Example: Traffic Deaths and Alcohol Taxes

We estimate simple regressions using data for years 1982 and 1988 that model the relationship between beer tax (adjusted for 1988 dollars) and the traffic fatality rate, measured as the number of fatalities per $10,000$ inhabitants.

Define the fatality rate:

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

Subset the data to the years of interest:

```
df1982 <- subset(df, year == "1982")
df1988 <- subset(df, year == "1988")
```

Estimate simple regression models using 1982 and 1988 data:

```
m1982 <- lm(fatality ~ beertax, data = df1982)
m1988 <- lm(fatality ~ beertax, data = df1988)
```

Display regression results with robust standard errors:

```
coeftest(m1982, vcov. = vcovHC, type = "HC1")
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.010      0.150   13.44   <2e-16 ***
## beertax        0.148      0.133    1.12     0.27
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coeftest(m1988, vcov. = vcovHC, type = "HC1")
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.859      0.115   16.22   <2e-16 ***
## beertax        0.439      0.128    3.43   0.0013 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

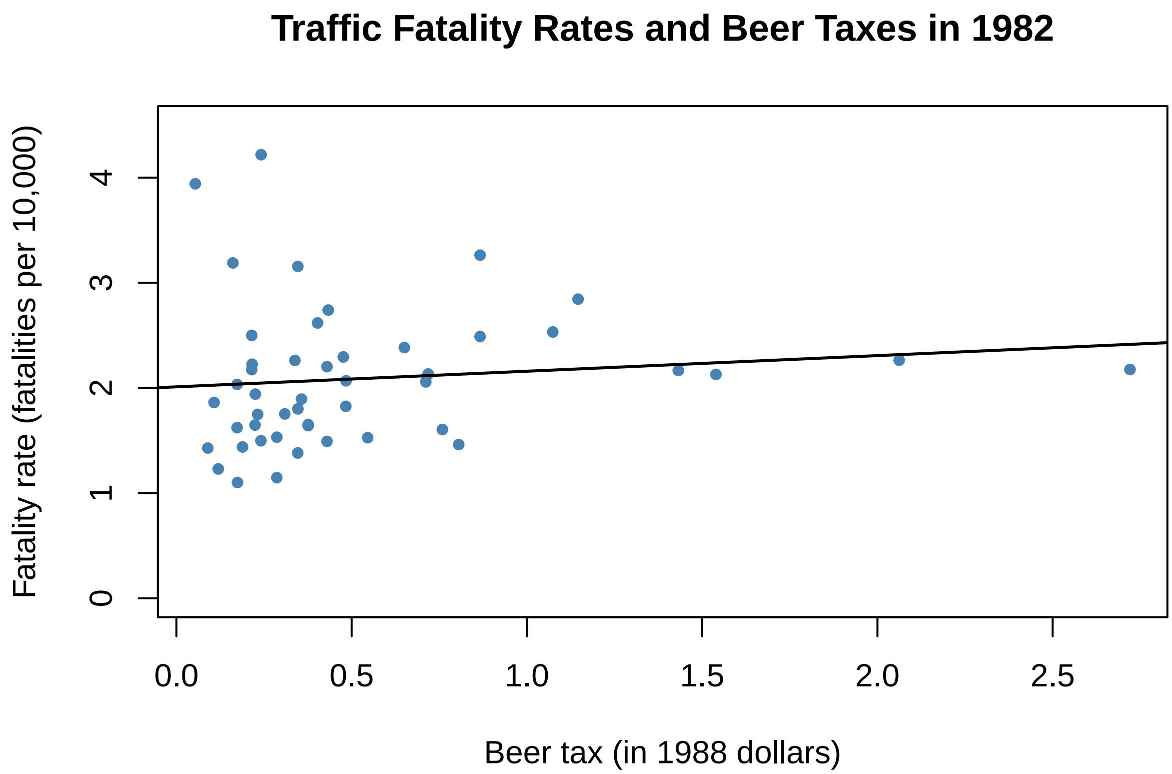
The estimated regression functions are

$$\widehat{FatalityRate} = \underset{(0.15)}{2.01} + \underset{(0.13)}{0.15}\, BeerTax \quad \text{(1982 data)},$$

$$\widehat{FatalityRate} = \underset{(0.11)}{1.86} + \underset{(0.13)}{0.44}\, BeerTax \quad \text{(1988 data)}.$$

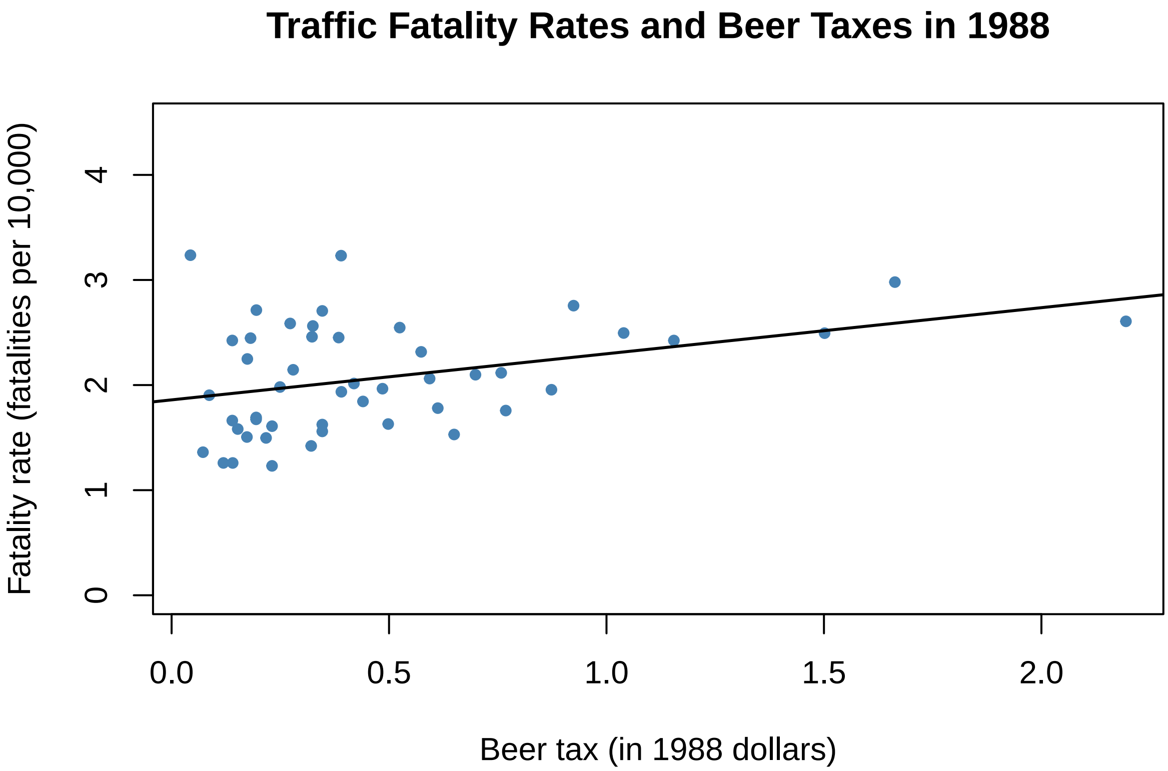
Plot observations and add the estimated regression line for 1982:

```
plot(x = df1982$beertax,
     y = df1982$fatality,
     xlab = "Beer tax (in 1988 dollars)",
     ylab = "Fatality rate (fatalities per 10,000)",
     main = "Traffic Fatality Rates and Beer Taxes in 1982",
     ylim = c(0, 4.5),
     pch = 20,
     col = "steelblue")

abline(m1982, lwd = 1.5)
```



Plot observations and add estimated regression line for 1988:

```
plot(x = df1988$beertax,
     y = df1988$fatality,
     xlab = "Beer tax (in 1988 dollars)",
     ylab = "Fatality rate (fatalities per 10,000)",
     main = "Traffic Fatality Rates and Beer Taxes in 1988",
     ylim = c(0, 4.5),
     pch = 20,
     col = "steelblue")

abline(m1988, lwd = 1.5)
```



The regression results indicate a positive relationship between the beer tax and the fatality rate for both years. The estimated coefficient on beer tax for the 1988 data is almost three times as large as for the 1988 dataset. This is contrary to our expectations: alcohol taxes are supposed to *lower* the rate of traffic fatalities. This apparent paradox could be due to omitted variable bias, since neither model includes covariates, e.g., economic conditions. A multiple regression analysis with suitable control variables could help address this problem. However, it cannot deal with omitted *unobservable* factors that differ from state to state while remaining constant over time, e.g. attitudes towards drunk driving. The next section uses panel data to hold such factors constant.

Panel Data with Two Time Periods: "Before and After" Comparisons

Suppose there are only $T = 2$ time periods $t = 1982, 1988$. This allows us to analyze differences in changes of the the fatality rate from year 1982 to 1988. Consider the population regression model:

$$FatalityRate_{it} = \beta_0 + \beta_1 BeerTax_{it} + \beta_2 Z_i + u_{it}$$

where the Z_i are state specific characteristics that differ between states but are *constant over time*. For $t = 1982$ and $t = 1988$ we have

$$\begin{aligned} FatalityRate_{i1982} &= \beta_0 + \beta_1 BeerTax_{i1982} + \beta_2 Z_i + u_{i1982}, \\ FatalityRate_{i1988} &= \beta_0 + \beta_1 BeerTax_{i1988} + \beta_2 Z_i + u_{i1988}. \end{aligned}$$

We can eliminate the Z_i by regressing the difference in the fatality rate between 1988 and 1982 on the difference in beer tax between those years:

$$FatalityRate_{i1988} - FatalityRate_{i1982} = \beta_1 (BeerTax_{i1988} - BeerTax_{i1982}) + u_{i1988} - u_{i1982}$$

This regression model yields an estimate for β_1 robust a possible bias due to omission of the Z_i, since these influences are eliminated from the model. Next we use use `R` to estimate a regression based on the differenced data and plot the estimated regression function.

Compute the differences:

```r
diff.fatality <- df1988$fatality - df1982$fatality
diff.beertax <- df1988$beertax - df1982$beertax
```

Estimate a regression using differenced data:

```r
m.diff <- lm(diff.fatality ~ diff.beertax)
coeftest(m.diff, vcov = vcovHC, type = "HC1")
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.0720     0.0654   -1.10   0.2761
## diff.beertax  -1.0410     0.3550   -2.93   0.0052 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

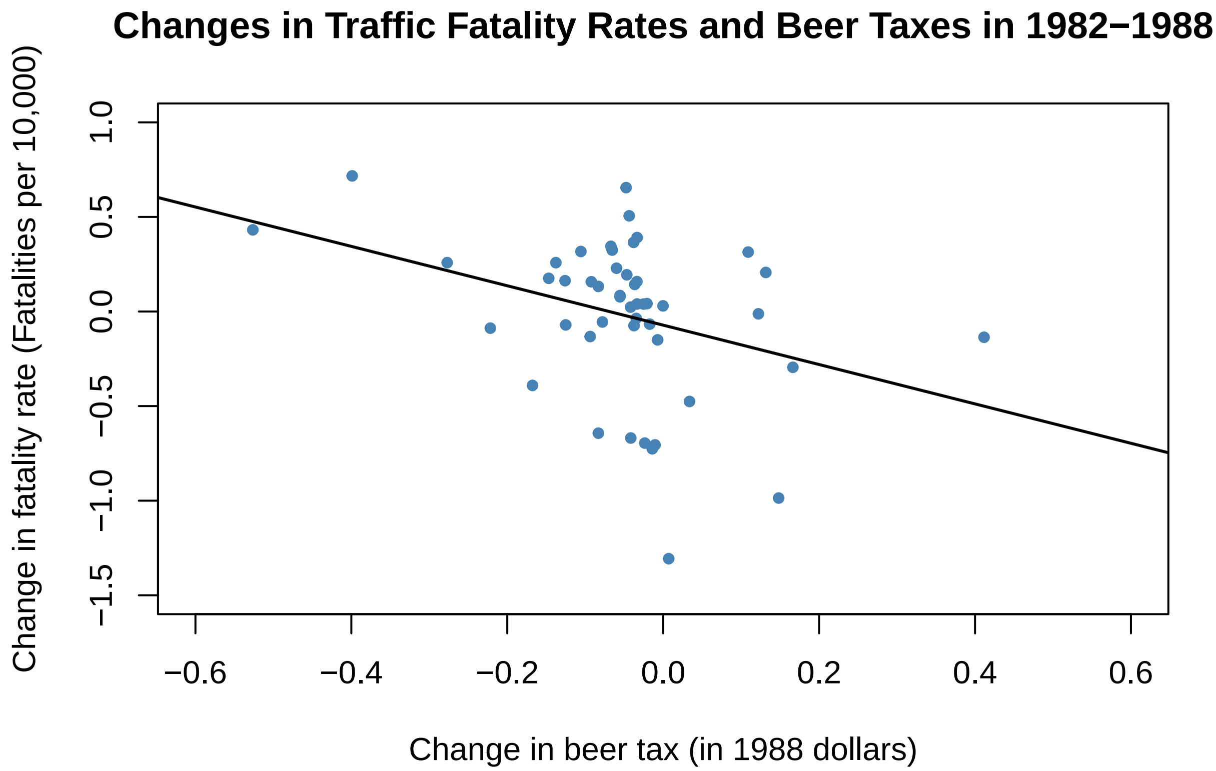
Including the intercept allows for a change in the mean fatality rate in the time between 1982 and 1988 in the absence of a change in the beer tax.

We obtain the OLS estimated regression function

$$FatalityRate_{i1988} \widehat{-} FatalityRate_{i1982} = -\underset{(0.065)}{0.072} - \underset{(0.36)}{1.04} \,(BeerTax_{i1988} - BeerTax_{i1982}).$$

Plot the differenced data:

```r
plot(x = diff.beertax,
     y = diff.fatality,
     xlab = "Change in beer tax (in 1988 dollars)",
     ylab = "Change in fatality rate (Fatalities per 10,000)",
     main = "Changes in Traffic Fatality Rates and Beer Taxes in 1982-1988",
     xlim = c(-0.6, 0.6),
     ylim = c(-1.5, 1),
     pch = 20,
     col = "steelblue")
# add the regression line to plot
abline(m.diff, lwd = 1.5)
```



The estimated coefficient on the beer tax is now negative and significantly different from zero at the 5% significance level. The interpretation is that raising the beer tax by \$1 causes traffic fatalities to decrease by 1.04 per $10,000$ people. This is quite large as the average fatality rate is approximately 2 persons per $10,000$ people.

Compute mean fatality rate over all states for all time periods:

```
mean(df$fatality)
```

```
## [1] 2.0404
```

Again this outcome is likely to be a consequence of omitting factors in the single-year regression that influence the fatality rate and are correlated with the beer tax *and* change over time. We need to control for such factors before drawing conclusions about the effect of a raise in beer taxes.

The Before/After comparison discards information for years 1983 to 1987. A method that allows to use data for more than $T = 2$ time periods and enables us to add control variables is the fixed effects regression approach.

Fixed Effects Regression

Consider the panel regression model

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + u_{it}$$

where the Z_i are unobserved time-invariant heterogeneities across the entities $i = 1, \dots, n$. We estimate β_1, the effect on Y_i of a change in X_i holding constant Z_i. Let $\alpha_i = \beta_0 + \beta_2 Z_i$. We obtain the model

$$Y_{it} = \alpha_i + \beta_1 X_{it} + u_{it} (\#eq : femodel). \tag{1}$$

Having individual specific intercepts α_i, $i = 1, \dots, n$, where each of these can be understood as the fixed effect of entity i, this model is called the *fixed effects model*. The variation in the α_i, $i = 1, \dots, n$ comes from the Z_i.

@ref(eq:femodel) can be rewritten as a regression model containing $n-1$ dummy regressors and a constant:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D2_i + \gamma_3 D3_i + \cdots + \gamma_n Dn_i + u_{it} (\#eq:drmodel). \tag{2}$$

Model @ref(eq:drmodel) has n different intercepts — one for every entity. @ref(eq:femodel) and @ref(eq:drmodel) are equivalent representations of the fixed effects model.

The fixed effects model can be generalized to contain more than just one determinant of Y that is correlated with X and changes over time.

Estimation and Inference

Software packages use a so-called "entity-demeaned" OLS algorithm which is computationally more efficient than estimating regression models with $k+n$ regressors as needed for models @ref(eq:gfemodel) and @ref(eq:gdrmodel).

Taking averages on both sides of @ref(eq:femodel) we obtain

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} Y_{it} &= \beta_1 \frac{1}{n}\sum_{i=1}^{n} X_{it} + \frac{1}{n}\sum_{i=1}^{n} a_i + \frac{1}{n}\sum_{i=1}^{n} u_{it} \\ \overline{Y} &= \beta_1 \overline{X}_i + \alpha_i + \overline{u}_i.\end{aligned}$$

Subtraction from @ref(eq:femodel) yields

$$\begin{aligned}Y_{it} - \overline{Y}_i &= \beta_1 (X_{it} - \overline{X}_i) + (u_{it} - \overline{u}_i) \\ \tilde{Y}_{it} &= \beta_1 \tilde{X}_{it} + \tilde{u}_{it}.\end{aligned} (\#eq:edols) \tag{3}$$

In this model, the OLS estimate of the parameter of interest β_1 is equal to the estimate obtained using @ref(eq:drmodel) — without the need to estimate $n-1$ dummies and an intercept.

There are two ways of estimating β_1 in the fixed effects regression:

1. OLS of the dummy regression model as shown in @ref(eq:drmodel)
2. OLS using the entity demeaned data as in @ref(eq:edols)

Provided the fixed effects regression assumptions for causal inference hold, the sampling distribution of the OLS estimator in the fixed effects regression model is normal in large samples. We now estimate a fixed effects model and report heteroskedasticity-robust standard errors.

Application to Traffic Deaths

The simple fixed effects model to estimate the relation between traffic fatality rates and the beer taxes includes 48 binary regressors — one for each federal state:

$$FatalityRate_{it} = \beta_1 BeerTax_{it} + StateFixedEffects + u_{it}, (\#eq:fatsemod) \tag{4}$$

The function $lm()$ can be used to estimate the slope coefficient β_1:

```
m.fe <- lm(fatality ~ beertax + state - 1, data=df)
summary(m.fe)
```

```
##
## Call:
## lm(formula = fatality ~ beertax + state - 1, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -0.5870 -0.0828 -0.0013  0.0795  0.8978
##
## Coefficients:
##         Estimate Std. Error t value Pr(>|t|)
## beertax  -0.6559     0.1878   -3.49  0.00056 ***
## stateal   3.4776     0.3134   11.10  < 2e-16 ***
## stateaz   2.9099     0.0925   31.45  < 2e-16 ***
## statear   2.8227     0.1321   21.36  < 2e-16 ***
## stateca   1.9682     0.0740   26.59  < 2e-16 ***
## stateco   1.9933     0.0804   24.80  < 2e-16 ***
## statect   1.6154     0.0839   19.25  < 2e-16 ***
## statede   2.1700     0.0775   28.02  < 2e-16 ***
## statefl   3.2095     0.2215   14.49  < 2e-16 ***
## statega   4.0022     0.4640    8.62  4.4e-16 ***
## stateid   2.8086     0.0988   28.44  < 2e-16 ***
## stateil   1.5160     0.0785   19.32  < 2e-16 ***
## statein   2.0161     0.0887   22.74  < 2e-16 ***
## stateia   1.9337     0.1022   18.92  < 2e-16 ***
## stateks   2.2544     0.1086   20.75  < 2e-16 ***
## stateky   2.2601     0.0805   28.09  < 2e-16 ***
## statela   2.6305     0.1627   16.17  < 2e-16 ***
## stateme   2.3697     0.1601   14.80  < 2e-16 ***
## statemd   1.7712     0.0825   21.48  < 2e-16 ***
## statema   1.3679     0.0865   15.82  < 2e-16 ***
## statemi   1.9931     0.1166   17.09  < 2e-16 ***
## statemn   1.5804     0.0936   16.88  < 2e-16 ***
## statems   3.4486     0.2094   16.47  < 2e-16 ***
## statemo   2.1814     0.0925   23.58  < 2e-16 ***
## statemt   3.1172     0.0944   33.02  < 2e-16 ***
## statene   1.9555     0.1055   18.53  < 2e-16 ***
## statenv   2.8769     0.0811   35.49  < 2e-16 ***
## statenh   2.2232     0.1411   15.75  < 2e-16 ***
## statenj   1.3719     0.0733   18.71  < 2e-16 ***
## statenm   3.9040     0.1015   38.45  < 2e-16 ***
## stateny   1.2910     0.0756   17.07  < 2e-16 ***
## statenc   3.1872     0.2517   12.66  < 2e-16 ***
## statend   1.8542     0.1019   18.19  < 2e-16 ***
## stateoh   1.8032     0.1019   17.69  < 2e-16 ***
## stateok   2.9326     0.1843   15.91  < 2e-16 ***
## stateor   2.3096     0.0812   28.45  < 2e-16 ***
## statepa   1.7102     0.0865   19.78  < 2e-16 ***
## stateri   1.2126     0.0775   15.64  < 2e-16 ***
## statesc   4.0348     0.3548   11.37  < 2e-16 ***
## statesd   2.4739     0.1412   17.52  < 2e-16 ***
## statetn   2.6020     0.0916   28.40  < 2e-16 ***
## statetx   2.5602     0.1085   23.59  < 2e-16 ***
## stateut   2.3137     0.1545   14.97  < 2e-16 ***
## statevt   2.5116     0.1397   17.98  < 2e-16 ***
## stateva   2.1874     0.1466   14.92  < 2e-16 ***
## statewa   1.8181     0.0823   22.08  < 2e-16 ***
## statewv   2.5809     0.1077   23.97  < 2e-16 ***
## statewi   1.7184     0.0775   22.18  < 2e-16 ***
## statewy   3.2491     0.0723   44.92  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.19 on 287 degrees of freedom
## Multiple R-squared:  0.993,  Adjusted R-squared:  0.992
## F-statistic:  848 on 49 and 287 DF,  p-value: <2e-16
```

It is also possible to estimate β_1 by applying OLS to the demeaned data, that is, to run the regression

$$Fatal\widetilde{ity}Rate = \beta_1 Be\widetilde{er}Tax_{it} + u_{it}.$$

To compute group averages, we can use the function *ave*. We first compute state-specific averages of the fatality rate and the beer tax and then run the regression on the de-meaned data:

```
df.demeaned <- with(df,
            data.frame(fatality = fatality - ave(fatality, state),
            beertax = beertax - ave(beertax, state)))
m.demean <- lm(fatality ~ beertax - 1, data=df.demeaned)
summary(m.demean)
```

```
##
## Call:
## lm(formula = fatality ~ beertax - 1, data = df.demeaned)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5870 -0.0828 -0.0013  0.0795  0.8978
##
## Coefficients:
##         Estimate Std. Error t value Pr(>|t|)
## beertax   -0.656      0.174   -3.77  0.00019 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.176 on 335 degrees of freedom
## Multiple R-squared:  0.0407, Adjusted R-squared:  0.0379
## F-statistic: 14.2 on 1 and 335 DF,  p-value: 0.000191
```

An alternative to *lm*() is the *plm*() function from the `plm` package. In addition to the regression formula and the data used, *plm*() requires a vector of names of entity and time variables passed to the argument *index*. The ID variable for entity effects is named *state* and the ID variable for time effects is *year*. To estimate a fixed effects model, set $model = "within"$. The function *coeftest*() can then be used to compute robust standard errors.

Estimate the fixed effects regression with *plm*():

```
library(plm)
```

```
##
## Attaching package: 'plm'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, lag, lead
```

```
plm(fatality ~ beertax,
    data = df,
    index = c("state", "year"),
```

```r
    model = "within") -> m.plm
coeftest(m.plm, vcov. = vcovHC, type="HC1")
```

```
##
## t test of coefficients:
##
##         Estimate Std. Error t value Pr(>|t|)
## beertax   -0.656      0.289   -2.27    0.024 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimated coefficient is again -0.6559. Note that $plm()$ uses the entity-demeaned OLS algorithm and thus does not report dummy coefficients. The estimated regression function is:

$$\widehat{FatalityRate} = -\underset{(0.29)}{0.66}\,BeerTax + StateFixedEffects.(\#eq:efemod) \tag{5}$$

The coefficient on $BeerTax$ is negative and significant. The interpretation is that the estimated reduction in traffic fatalities due to an increase in the real beer tax by \$1 is 0.66 per $10,000$ people, which is still pretty high. Although including state fixed effects eliminates the risk of a bias due to omitted factors that vary across states but not over time, we suspect that there are other omitted variables that vary over time and thus cause a bias.

Regression with Time Fixed Effects

Controlling for variables that are constant across entities but vary over time can be done by including time fixed effects. If there are *only* time fixed effects, the fixed effects regression model becomes

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \delta_2 B2_t + \cdots + \delta_T BT_t + u_{it}$$

where only $T-1$ dummies are included ($B1$ is omitted) since the model includes an intercept. This model eliminates omitted variable bias caused by excluding unobserved variables that evolve over time but are constant across entities.

In some applications it is meaningful to include both entity and time fixed effects. The *entity and time fixed effects model* is:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D2_i + \cdots + \gamma_n DT_i + \delta_2 B2_t + \cdots + \delta_T BT_t + u_{it}$$

The combined model can be used to eliminate 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.

Estimate the combined entity and time fixed effects model of the relation between fatalities and beer tax:

$$FatalityRate_{it} = \beta_1 BeerTax_{it} + StateEffects + TimeFixedEffects + u_{it}$$

using both $lm()$ and $plm()$. It is straightforward to estimate this regression with $lm()$ since it is just an extension of @ref(eq:fatsemod) so we only have to adjust the $formula$ argument by adding the additional regressor *year* for time fixed effects. In our call of $plm()$ we set another argument $effect = "twoways"$ for inclusion of entity *and* time dummies.

Estimate a regression model with both time and entity fixed effects

with $lm()$:

```r
lm(fatality ~ beertax + state + year - 1, data=df) -> m.fete.lm
summary(m.fete.lm)
```

```
##
## Call:
## lm(formula = fatality ~ beertax + state + year - 1, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5956 -0.0810  0.0014  0.0823  0.8388
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## beertax    -0.6400     0.1974   -3.24   0.0013 **
## stateal     3.5114     0.3325   10.56  < 2e-16 ***
## stateaz     2.9645     0.0993   29.85  < 2e-16 ***
## statear     2.8728     0.1416   20.29  < 2e-16 ***
## stateca     2.0262     0.0786   25.79  < 2e-16 ***
## stateco     2.0498     0.0859   23.85  < 2e-16 ***
## statect     1.6712     0.0899   18.59  < 2e-16 ***
## statede     2.2271     0.0826   26.95  < 2e-16 ***
## statefl     3.2513     0.2359   13.78  < 2e-16 ***
## statega     4.0230     0.4909    8.20  8.9e-15 ***
## stateid     2.8624     0.1061   26.99  < 2e-16 ***
## stateil     1.5729     0.0838   18.77  < 2e-16 ***
## statein     2.0712     0.0951   21.78  < 2e-16 ***
## stateia     1.9871     0.1098   18.10  < 2e-16 ***
## stateks     2.3071     0.1166   19.78  < 2e-16 ***
## stateky     2.3166     0.0860   26.92  < 2e-16 ***
## statela     2.6777     0.1739   15.40  < 2e-16 ***
## stateme     2.4171     0.1712   14.12  < 2e-16 ***
## statemd     1.8273     0.0883   20.70  < 2e-16 ***
## statema     1.4234     0.0927   15.35  < 2e-16 ***
## statemi     2.0449     0.1252   16.34  < 2e-16 ***
## statemn     1.6349     0.1005   16.27  < 2e-16 ***
## statems     3.4915     0.2231   15.65  < 2e-16 ***
## statemo     2.2360     0.0993   22.52  < 2e-16 ***
## statemt     3.1716     0.1014   31.29  < 2e-16 ***
## statene     2.0085     0.1133   17.73  < 2e-16 ***
## statenv     2.9332     0.0867   33.83  < 2e-16 ***
## statenh     2.2724     0.1512   15.03  < 2e-16 ***
## statenj     1.4302     0.0777   18.40  < 2e-16 ***
## statenm     3.9575     0.1090   36.30  < 2e-16 ***
## stateny     1.3485     0.0805   16.75  < 2e-16 ***
## statenc     3.2263     0.2677   12.05  < 2e-16 ***
## statend     1.9076     0.1095   17.43  < 2e-16 ***
## stateoh     1.8566     0.1095   16.96  < 2e-16 ***
## stateok     2.9778     0.1967   15.14  < 2e-16 ***
## stateor     2.3660     0.0868   27.24  < 2e-16 ***
## statepa     1.7656     0.0927   19.04  < 2e-16 ***
## stateri     1.2696     0.0827   15.35  < 2e-16 ***
## statesc     4.0650     0.3761   10.81  < 2e-16 ***
## statesd     2.5232     0.1512   16.68  < 2e-16 ***
## statetn     2.6567     0.0983   27.02  < 2e-16 ***
```

```
## statetx    2.6128     0.1165   22.42  < 2e-16 ***
## stateut    2.3617     0.1653   14.29  < 2e-16 ***
## statevt    2.5610     0.1497   17.11  < 2e-16 ***
## stateva    2.2362     0.1570   14.25  < 2e-16 ***
## statewa    1.8742     0.0881   21.27  < 2e-16 ***
## statewv    2.6336     0.1156   22.78  < 2e-16 ***
## statewi    1.7754     0.0826   21.49  < 2e-16 ***
## statewy    3.3079     0.0764   43.29  < 2e-16 ***
## year1983  -0.0799     0.0384   -2.08   0.0381 *
## year1984  -0.0724     0.0384   -1.89   0.0600 .
## year1985  -0.1240     0.0384   -3.23   0.0014 **
## year1986  -0.0379     0.0386   -0.98   0.3273
## year1987  -0.0509     0.0390   -1.31   0.1926
## year1988  -0.0518     0.0396   -1.31   0.1921
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.188 on 281 degrees of freedom
## Multiple R-squared:  0.993,  Adjusted R-squared:  0.992
## F-statistic:  772 on 55 and 281 DF,  p-value: <2e-16
```

The $lm()$ functions converts factors into dummies automatically. Since we exclude the intercept by adding -1 to the right-hand side of the regression formula, $lm()$ estimates coefficients for $n + (T - 1) = 48 + 6 = 54$ binary variables (6 year dummies and 48 state dummies).

With $plm()$:

```r
plm(fatality ~ beertax,
    data = df,
    index = c("state", "year"),
    model = "within",
    effect = "twoways") -> m.fete.plm
# check class
class(m.fete.plm)
```

```
## [1] "plm"        "panelmodel"
```

```r
coeftest(m.fete.plm, vcov = vcovHC, type = "HC1")
```

```
##
## t test of coefficients:
##
##         Estimate Std. Error t value Pr(>|t|)
## beertax    -0.64       0.35   -1.83    0.069 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$plm()$ only reports the estimated coefficient on $BeerTax$.

The estimated regression function is

$$\widehat{FatalityRate} = -\underset{(0.35)}{0.64}\, BeerTax + StateEffects + TimeFixedEffects.(\#eq:cbnfemod) \tag{6}$$

The result -0.66 is close to the estimated coefficient for the regression model including only entity fixed effects. Unsurprisingly, the coefficient is less precisely estimated but significantly different from zero at 10%.

From the results in @ref(eq:efemod) and @ref(eq:cbnfemod), we conclude that the estimated relationship between traffic fatalities and the real beer tax is not affected by omitted variable bias due to factors that are constant over time.

Standard Errors for Fixed Effects Regression

If there is evidence of both heteroskedasticity *and* autocorrelation *heteroskedasticity and autocorrelation-consistent (HAC) standard errors* need to be used. *Clustered standard errors* allow for heteroskedasticity and autocorrelated errors within an entity but *not* correlated across entities.

Clustered standard errors can be estimated with $coeftest()$ in conjunction with $vcovHC()$ from the package *sandwich*. Conveniently, $vcovHC()$ recognizes panel model objects (objects of class *plm*) and computes clustered standard errors by default.

It is crucial to use clustered standard errors in empirical applications of fixed effects models. To see this, consider the entity and time fixed effects model for fatalities. By default, $coeftest()$ uses robust standard errors that are only valid in the absence of autocorrelated errors.

Heteroskedasticity-robust standard errors (but not robust to autocorrelation):

```
coeftest(m.fete.lm, vcov = vcovHC, type = "HC1")[1,]
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
##  -0.639980   0.254715  -2.512535   0.012547
```

Clustered standard errors (robust to both heteroskedasticity and autocorrelation):

```
coeftest(m.fete.plm, vcov = vcovHC, type = "HC1")
```

```
##
## t test of coefficients:
##
##         Estimate Std. Error t value Pr(>|t|)
## beertax    -0.64       0.35   -1.83    0.069 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The outcomes differ: imposing no autocorrelation we obtain a standard error of 0.25 which implies significance of $\hat{\beta}_1$, the coefficient on $BeerTax$ at the level of 5%. By contrast, using the clustered standard error 0.35 leads to acceptance of the hypothesis $H_0 : \beta_1 = 0$ at the same level.

Drunk Driving Laws and Traffic Deaths

There are two major sources of omitted variable bias that are not accounted for by the models considered so far: economic conditions and driving laws. The dataset contains state-specific legal drinking age *drinkage*, punishment (*jail*, *service*) and various economic indicators like unemployment rate (*unemp*) and per capita income (*income*). We will use these covariates to extend the preceding analysis.

These covariates are defined as follows:

- *unemp*: a numeric variable stating the state specific unemployment rate.
- $log(income)$: the logarithm of real per capita income (in prices of 1988).
- *miles*: the state average miles per driver.
- *drinkage*: the state specify minimum legal drinking age.
- *drinkagc*: a discretized version of *drinkage* that classifies states into four categories of minimal drinking age; 18, 19, 20, 21 and older. `R` denotes this as $[18, 19)$, $[19, 20)$, $[20, 21)$ and $[21, 22]$. These categories are included as dummy regressors where $[21, 22]$ is chosen as the reference category.
- *punish*: a dummy variable with levels *yes* and *no* that measures if drunk driving is severely punished by mandatory jail time or mandatory community service (first conviction).

Define the variables according to the regression results presented in Table 10.1 of the book.

Discretize the minimum legal drinking age:

```
df$drinkage.factor <- cut(df$drinkage,
                    breaks = 18:22,
                    include.lowest = TRUE,
                    right = FALSE)
```

Set minimum drinking age [21, 22] to be the baseline level:

```
df$drinkage.factor <- relevel(df$drinkage.factor, "[21,22]")
```

Dummy for mandadory jail or community service

```
df$punish <- with(df, factor(jail == "yes" | service == "yes",
                                            labels = c("no", "yes")))
```

All variables for 1982 and 1988:

```
df.1982.1988 <- df[with(df, year == 1982 | year == 1988), ]
```

Estimate all seven models using *plm*().

```
m1 <- lm(fatality ~ beertax, data = df)

m2 <- plm(fatality ~ beertax + state, data = df)

m3 <- plm(fatality ~ beertax + state + year,
                       index = c("state","year"),
                       model = "within",
                       effect = "twoways",
                       data = df)

m4 <- plm(fatality ~ beertax + state + year + drinkage.factor
                       + punish + miles + unemp + log(income),
                       index = c("state", "year"),
                       model = "within",
                       effect = "twoways",
                       data = df)

m5 <- plm(fatality ~ beertax + state + year + drinkage.factor
                       + punish + miles,
                       index = c("state", "year"),
                       model = "within",
                       effect = "twoways",
                       data = df)

m6 <- plm(fatality ~ beertax + year + drinkage
                       + punish + miles + unemp + log(income),
                       index = c("state", "year"),
                       model = "within",
                       effect = "twoways",
                       data = df)

m7 <- plm(fatality ~ beertax + state + year + drinkage.factor
                       + punish + miles + unemp + log(income),
                       index = c("state", "year"),
                       model = "within",
```

```r
                        effect = "twoways",
                        data = df.1982.1988)
```

Use *stargazer*() to generate a table of the results. Clustered standard errors are stored in a list and passed to the *se* argument to generate the table:

```r
# library(stargazer)
se.robust <- list(sqrt(diag(vcovHC(m1, type = "HC1"))),
                  sqrt(diag(vcovHC(m2, type = "HC1"))),
                  sqrt(diag(vcovHC(m3, type = "HC1"))),
                  sqrt(diag(vcovHC(m4, type = "HC1"))),
                  sqrt(diag(vcovHC(m5, type = "HC1"))),
                  sqrt(diag(vcovHC(m6, type = "HC1"))),
                  sqrt(diag(vcovHC(m7, type = "HC1"))))
```

The *stargazer*() function can generate tables suitable for different formats, including "html" and "pdf". To create LaTeX output, set *type* = "*latex*". To create HTML output, set *type* = "*html*". To automate the process, we first save the output type with `rmd.type <- knitr::opts_knit$get("rmarkdown.pandoc.to")` and then pass it to *stargazer*(), to ensure that the appropriate table is generated when knitting to "html" and "pdf".

```r
rmd.type <- knitr::opts_knit$get("rmarkdown.pandoc.to")
    # returns "html" when knitting to "html"
    # returns "latex" when knitting to "pdf"
stargazer(m1, m2, m3, m4, m5, m6, m7,
          type = rmd.type,
          se = se.robust,
          header = FALSE,
          column.sep.width = "-20pt")
```

The above table is too wide to display properly in a standard PDF document. To fit the table in PDF format, we select the `"sidewaystable"` option and squeeze the inter-column space by setting `column.sep.width` to a negative value. We also fix the column label and, to clean up the output, remove the row of F statistics. And we set the style to the *Quarterly Journal of Economics* with `style="qje"`.

```r
stargazer(m1, m2, m3, m4, m5, m6, m7,
          type = rmd.type,
          se = se.robust,
          style = "qje",
          float.env = "sidewaystable",
          column.labels = c('OLS','','','Linear Panel Regression'),
          omit.stat = "f",
          title = "Linear Panel Regression Models of Traffic Fatalities due to Drunk Driving",
          header = FALSE,
          model.names = FALSE,
          model.numbers = TRUE,
          object.names=FALSE,
          digits = 3,
          column.sep.width = "0pt")
```

While columns (2) and (3) sum up the results of @ref(eq:efemod) and @ref(eq:cbnfemod), column (1) presents an estimate of the coefficient of interest in the basic OLS regression without fixed effects. The estimate of the coefficient on beer tax is *positive* and likely to be biased upwards. The model fit is poor ($\bar{R}^2 = 0.091$). The sign of the estimate changes as we extend the model by both entity and time fixed effects in models (2) and (3). Furthermore $\bar{R}^2$ increases substantially as fixed effects are included in the model equation. The magnitudes of both estimates are likely too large.

Table 1:

	Dependent variable:				
	fatality				
	OLS			*panel linear*	
	(1)	(2)	(3)	(4)	(5)
beertax	0.365***	−0.656**	−0.640*	−0.445	−0.690**
	(0.053)	(0.289)	(0.350)	(0.291)	(0.345)
drinkage.factor[18,19)				0.028	−0.010
				(0.068)	(0.081)
drinkage.factor[19,20)				−0.018	−0.076
				(0.049)	(0.066)
drinkage.factor[20,21)				0.032	−0.100*
				(0.050)	(0.055)
drinkage					
punishyes				0.038	0.085
				(0.101)	(0.109)
miles				0.00001	0.00002*
				(0.00001)	(0.00001)
unemp				−0.063***	
				(0.013)	
log(income)				1.816***	
				(0.624)	
Constant	1.853***				
	(0.047)				
Observations	336	336	336	335	335
R^2	0.093	0.041	0.036	0.360	0.066
Adjusted R^2	0.091	−0.120	−0.149	0.217	−0.134
Residual Std. Error	0.544 (df = 334)				
F Statistic	34.394*** (df = 1; 334)	12.190*** (df = 1; 287)	10.513*** (df = 1; 281)	19.194*** (df = 8; 273)	3.252*** (df = 6; 2

Note:

Table 2: Linear Panel Regression Models of Traffic Fatalities due to Drunk Driving

	fatality						
	OLS	Linear Panel Regression					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
beertax	0.365***	−0.656**	−0.640*	−0.445	−0.690**	−0.456	−0.926***
	(0.053)	(0.289)	(0.350)	(0.291)	(0.345)	(0.301)	(0.337)
drinkage.factor[18,19)				0.028	−0.010		0.037
				(0.068)	(0.081)		(0.101)
drinkage.factor[19,20)				−0.018	−0.076		−0.065
				(0.049)	(0.066)		(0.097)
drinkage.factor[20,21)				0.032	−0.100*		−0.113
				(0.050)	(0.055)		(0.123)
drinkage						−0.002	
						(0.021)	
punishyes				0.038	0.085	0.039	0.089
				(0.101)	(0.109)	(0.101)	(0.161)
miles				0.00001	0.00002*	0.00001	0.0001***
				(0.00001)	(0.00001)	(0.00001)	(0.00005)
unemp				−0.063***		−0.063***	−0.091***
				(0.013)		(0.013)	(0.021)
log(income)				1.816***		1.786***	0.996
				(0.624)		(0.631)	(0.666)
Constant	1.853***						
	(0.047)						
N	336	336	336	335	335	335	95
R^2	0.093	0.041	0.036	0.360	0.066	0.357	0.659
Adjusted R^2	0.091	−0.120	−0.149	0.217	−0.134	0.219	0.157
Residual Std. Error	0.544 (df = 334)						

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

The model specifications (4) to (7) include covariates intended to capture the effect of economic conditions and the legal environment.

Consider (4) as the baseline specification. Four interesting results stand out:

1. Including the covariates does not lead to a major reduction of the estimated effect of the beer tax. The coefficient is not significantly different from zero at the level of 5% as the estimate is rather imprecise.
2. The minimum legal drinking age *does not* have an effect on traffic fatalities: none of the three dummy variables are significantly different from zero at typical levels of significance. The F-Test of the joint hypothesis that all three coefficients are zero cannot reject the null hypothesis of no joint effect.

Test if legal drinking age has no explanatory power

```
linearHypothesis(m4,
                 test = "F",
                 c("drinkage.factor[18,19)=0",
                   "drinkage.factor[19,20)=0",
                   "drinkage.factor[20,21)"),
                 vcov. = vcovHC, type = "HC1")
```

```
## Linear hypothesis test
##
## Hypothesis:
## drinkage.factor[18,19) = 0
## drinkage.factor[19,20) = 0
## drinkage.factor[20,21) = 0
##
## Model 1: restricted model
## Model 2: fatality ~ beertax + state + year + drinkage.factor + punish +
##     miles + unemp + log(income)
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df    F Pr(>F)
## 1    276
## 2    273  3 0.38   0.77
```

3. There is no evidence that punishment for first offenders has a deterring effects on drunk driving: The estimated coefficient is not significant at the 10% level.
4. The economic variables significantly explain traffic fatalities. The employment rate and per capita income are jointly significant at the level of 0.1%.

```
linearHypothesis(m4,
                 test = "F",
                 c("log(income)", "unemp"),
                 vcov. = vcovHC, type = "HC1")
```

```
## Linear hypothesis test
##
## Hypothesis:
## log(income) = 0
## unemp = 0
##
## Model 1: restricted model
## Model 2: fatality ~ beertax + state + year + drinkage.factor + punish +
##     miles + unemp + log(income)
```

```
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df    F  Pr(>F)
## 1    275
## 2    273  2 31.6 4.6e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model (5) omits controls for economic conditions. The coefficient on beer tax is sensitive to the inclusion of controls for economic conditions, suggesting that they should be included.

Model (6) shows that the legal drinking age has little explanatory power and that the coefficient of interest is not sensitive to changes in the functional form of the relation between drinking age and traffic fatalities.

Model (7) shows that reducing the amount of available information (using 95 observations for the period 1982 to 1988) inflates standard errors but does not lead to drastic changes in coefficient estimates.

Conclusion

There is no evidence that increasing punishment and increasing the minimum drinking age reduce traffic fatalities due to drunk driving. There is a negative effect of alcohol taxes on traffic fatalities, but it is imprecisely estimated and cannot be interpreted as the causal effect of interest. The main drawback of this analysis is that there may be omitted variables that differ across states *and* change over time: This potential bias is not eliminated by controlling for entity specific and time invariant unobservables.

Instrumental variables regression can provide a way around the omitted variable bias where fixed effect panel regression techniques cannot.