计量经济学

第九讲：面板数据回归

黄嘉平

工学博士 经济学博士
深圳大学中国经济特区研究中心 讲师

办公室　粤海校区汇文楼2613
E-mail　huangjp@szu.edu.cn
Website　https://huangjp.com

主要内容

- 面板数据
- 面板数据回归
    - “前后”比较
    - **固定效应回归**
    - 时间固定效应回归
    - 固定效应回归假设与标准误
    - 固定效应与随机效应

面板数据

面板数据

Panel data

- **面板数据 (panel data)**，也称为**纵向数据 (longitudinal data)**，是指 n 个不同个体在 T 个不同时期上的观测数据。

TABLE 1.3 Selected Observations on Cigarette Sales, Prices, and Taxes, by State and Year for U.S. States, 1985–1995

Observation Number	State	Year	Cigarette Sales (packs per capita)	Average Price per Pack (including taxes)	Total Taxes (cigarette excise tax + sales tax)
1	Alabama	1985	116.5	$1.022	$0.333
2	Arkansas	1985	128.5	1.015	0.370
3	Arizona	1985	104.5	1.086	0.362
⋮	⋮	⋮	⋮	⋮	⋮
47	West Virginia	1985	112.8	1.089	0.382
48	Wyoming	1985	129.4	0.935	0.240
49	Alabama	1986	117.2	1.080	0.334
⋮	⋮	⋮	⋮	⋮	⋮
96	Wyoming	1986	127.8	1.007	0.240
97	Alabama	1987	115.8	1.135	0.335
⋮	⋮	⋮	⋮	⋮	⋮
528	Wyoming	1995	112.2	1.585	0.360

Cigarette 数据集包含美国48个大陆州在11年间的观测数据

- 如果数据集包含变量 X 和 Y 的观测值，该数据可以表示为

$$(X_{it}, Y_{it}), \quad i = 1,\ldots,n \text{ and } t = 1,\ldots,T$$

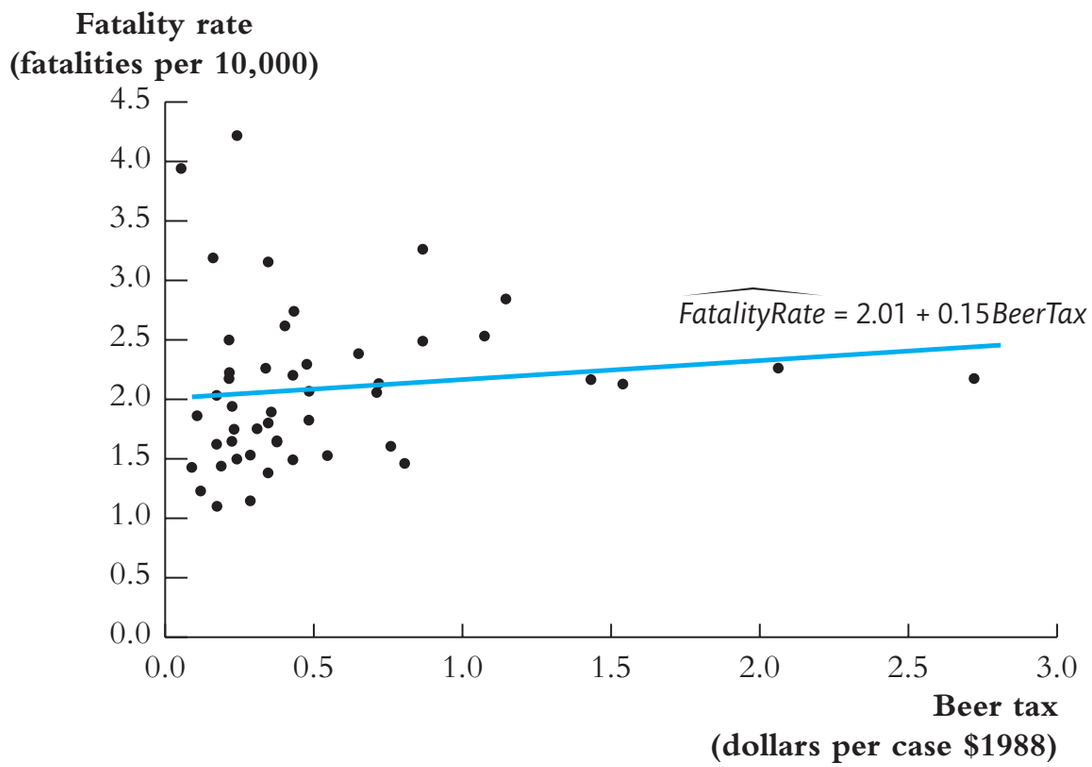
- **平衡面板（balanced panel）**指所有的观测值，即变量在每个个体和每一时期中都能被观测到。否则称为**非平衡面板（unbalanced panel）**。

美国州级交通事故死亡数据集

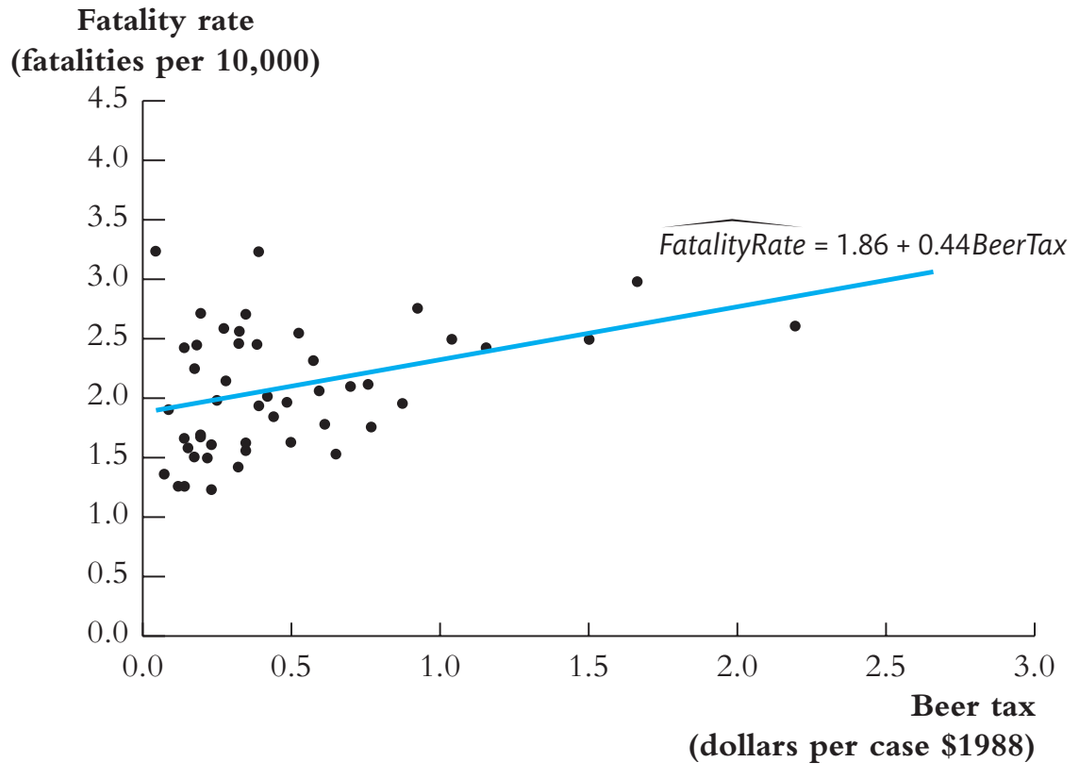
The U.S. state traffic fatality data set

- 数据文件：`fatality.xlsx`　说明文件：`fatality.docx`
- 包含美国 48 个本土州（不包含阿拉斯加和夏威夷）1982 至 1988 年间的交通事故及相关的平衡面板数据。
- 主要变量：
  - 交通事故死亡率：每州每年每一万人中死于交通事故的人数
  - 啤酒税：单位为美元/箱（1988年美元价值），可作为酒精税的代理变量
  - 法定饮酒年龄：分别为18岁、19岁、20岁的指示变量
  - 酒驾处罚：各州对首次酒驾犯罪的最低判决要求
  - 个人收入、失业率、人口等特征变量。

交通事故死亡率与酒精税



(a) 1982 data



(b) 1988 data

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

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

一元线性回归的结果是反常识的：啤酒税越高死亡率也越高

⇒ 存在遗漏变量偏差：道路设施状态、社会对饮酒的容忍度等

很难度量，却不随时间而变化 → 固定效应 OLS 回归

面板数据回归

“前后”比较

“Before and after” comparisons

- 这种方法适用于仅有两个时期的数据时：即 $T = 2$ 。
- 令 Z_i 表示决定第 i 个州死亡率的变量，但不随时间变化（所以省略的下角标 t ）。
- 考虑线性回归模型

$$\text{FatalityRate}_{it} = \beta_0 + \beta_1 \text{BeerTax}_{it} + \beta_2 Z_i + u_{it}$$

其中 u_{it} 为误差项，$i = 1,\ldots,n; t = 1,\ldots,T$ 。

“前后”比较

“Before and after” comparisons

- 假设我们只有 1982 年和 1988 年的数据，则有下面两个回归方程：

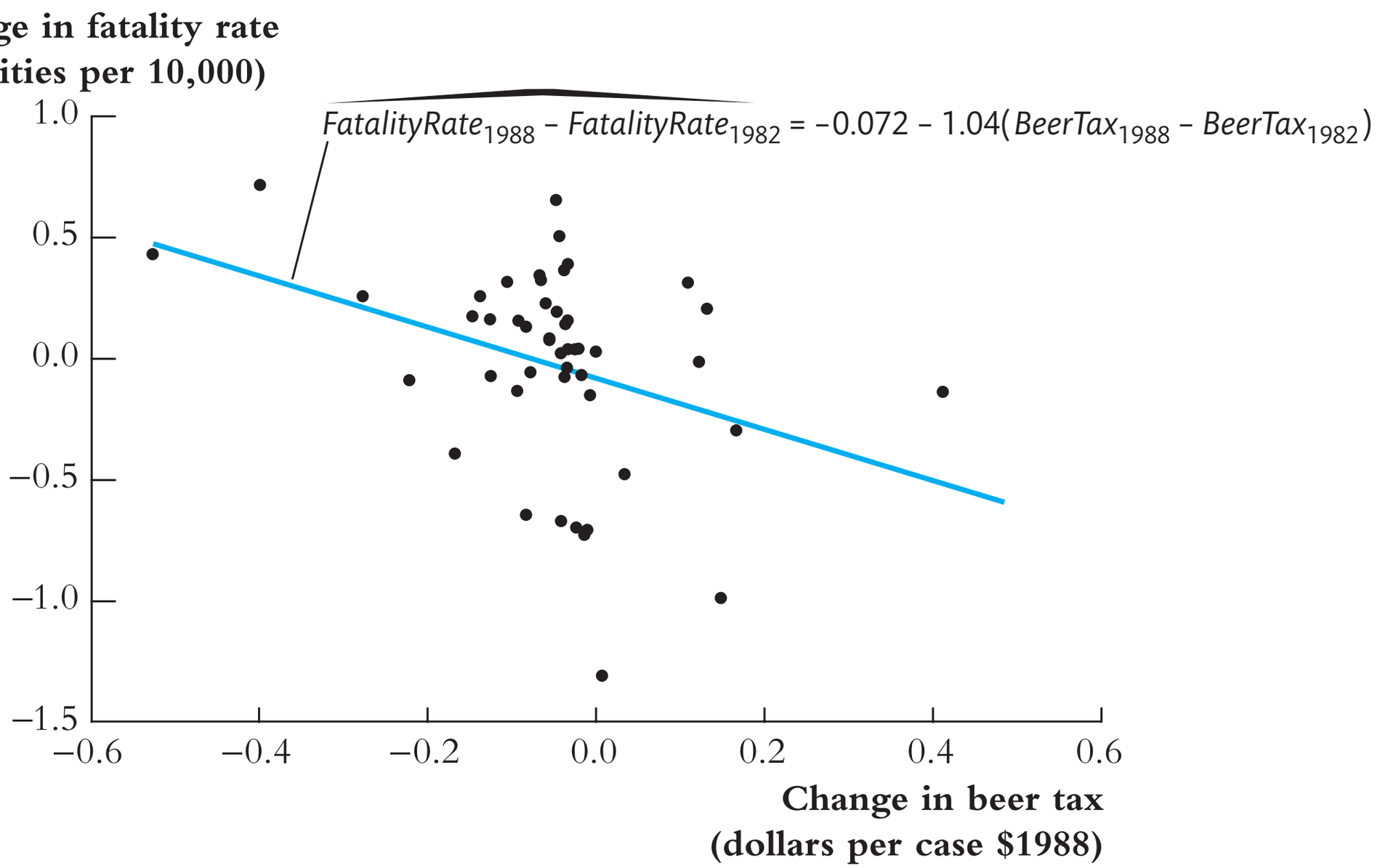
$$\text{FatalityRate}_{i1982} = \beta_0 + \beta_1 \text{BeerTax}_{i1982} + \beta_2 Z_i + u_{i1982}$$

$$\text{FatalityRate}_{i1988} = \beta_0 + \beta_1 \text{BeerTax}_{i1988} + \beta_2 Z_i + u_{i1988}$$

继而可得出

$$\begin{aligned}\text{FatalityRate}_{i1988} - \text{FatalityRate}_{i1982} \\ = \beta_1(\text{BeerTax}_{i1988} - \text{BeerTax}_{i1982}) + u_{i1988} - u_{i1982}\end{aligned}$$

这里消除了 Z_i 的效应



$$\overbrace{FatalityRate_{1988} - FatalityRate_{1982}} = \underset{(0.065)}{-0.072} - \underset{(0.36)}{1.04}(BeerTax_{1988} - BeerTax_{1982}).$$

其他变量的变化（如汽车安全性的提高）使死亡率下降了（−0.072）

啤酒税每上涨 1 美元/杯，死亡率减少 1.04 人/万人

练习：尝试在 gretl 中复制 (10.8) 式

- 定义交通事故死亡率

  ```
  series fatality = allmort / pop * 10000
  ```

  注：数据中 mrall 的计算方式应为 allmort / pop，并非书中用到的每万人死亡人数。

- 处理数据时的常用命令

  smpl — resets the sample range
  store — save data to a file (by default a .gdt file)
  open — opens a data file
  append — opens a data file and appends to the current dataset

- 也可以在 Excel 中事先处理好数据，再导入 gretl。

固定效应回归

Fixed effects regression

- 考虑线性回归模型

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

其中 Z_i 是随个体 i 变化但不随时间变化的不可观测变量，u_{it} 为误差项，$i = 1, \ldots, n; t = 1, \ldots, T$。我们想要估计 β_1，即固定不可观测的个体特征 Z 的情况下 X 对 Y 的效应。

- 以上模型可以理解为具有 n 个截距，每个截距对应一个个体 i。令 $\alpha_i = \beta_0 + \beta_2 Z_i$，则有可将以上模型改写为

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}$$

称为**固定效应回归模型（fixed effects regression model）**，其中 $\alpha_1, \ldots, \alpha_n$ 被称为**个体固定效应（entity/individual fixed effects）**。

二值变量设定

- 固定效应回归模型 $Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}$ 可以利用 $n-1$ 个二值变量等价地表示为

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D2_i + \gamma_3 D3_i + \cdots + \gamma_n Dn_i + u_{it}$$

其中 Dk_i , $k = 2,\ldots,n$ 是当 $i = k$ 时取值为 1 的二值变量。由此式我们可以估计系数 $\beta_0, \beta_1, \gamma_2, \gamma_3, \ldots, \gamma_n$ （以 $n \times T$ 个观测值估计 $n+1$ 个参数）。

- 个体固定效应为

$$\alpha_1 = \beta_0,$$
$$\alpha_i = \beta_0 + \gamma_i \quad \text{for } i \geq 2$$

包含多元回归变量的固定效用回归模型

- 固定效应回归模型为

$$Y_{it} = \beta_1 X_{1,it} + \cdots + \beta_k X_{k,it} + \alpha_i + u_{it},$$

$i = 1,\ldots,n;\ t = 1,\ldots,T$。

- 等价地，以上模型也可以用一个共同截距，独立变量 X，以及 $n-1$ 个二值变量表示：

$$\begin{aligned} Y_{it} =& \beta_0 + \beta_1 X_{1,it} + \cdots + \beta_k X_{k,it} \\ &+ \gamma_2 D2_i + \gamma_3 D3_i + \cdots + \gamma_n Dn_i + u_{it} \end{aligned}$$

估计和推断

- 一般固定效应回归模型的二值变量设定包含 $k+n$ 个参数需要估计。这在 n 较大时会使 OLS 回归的计算量增加，导致某些软件无法运行。

- “个体中心化”（entity-demeaned）OLS 算法

1. 计算变量的个体均值，并从每个变量中减去。以一元模型 $Y_{it}=\beta_1 X_{it}+\alpha_i+u_{it}$ 为例，令

$$\overline{Y}_i=\sum_{t=1}^{T}Y_{it}/T,\quad \overline{X}_i=\sum_{t=1}^{T}X_{it}/T,\quad \overline{u}_i=\sum_{t=1}^{T}u_{it}/T,$$

并定义 $\tilde{Y}_{it}=Y_{it}-\overline{Y}_i,\ \tilde{X}_{it}=X_{it}-\overline{X}_i,\ \tilde{u}_{it}=u_{it}-\overline{u}_i$。

2. 对 $\tilde{Y}_{it}=\beta_1\tilde{X}_{it}+\tilde{u}_{it}$ 进行 OLS 回归。

练习：用交通事故数据尝试各种回归方法

- 将数据设置为面板形式

```
setobs state year --panel-vars
```

entity var.　　time var.

- 二值变量法

```
genr unitdum  # create 48 unit(entity) dummies
ols fatality beertax du_* --robust
    # const is not needed
    # du_* indicates all unit dummies
```

- “个体中心化” 法

```
panel fatality const beertax --robust
    # const is needed
```

注：panel 命令默认使用“个体中心化”法估计个体固定效应模型，详见 gretl 帮助文件。

```
Model 1: Pooled OLS, using 336 observations
Included 48 cross-sectional units
Time-series length = 7
Dependent variable: fatality
Robust (HAC) standard errors

            coefficient   std. error      z       p-value
  ------------------------------------------------------------
  beertax     -0.655874    0.314848     -2.083   0.0372    **
  du_1         3.47763     0.511247      6.802   1.03e-11  ***
  du_2         2.90990     0.0979303    29.71    5.06e-194 ***
  du_3         2.82268     0.185941     15.18    4.76e-52  ***
  du_4         1.96816     0.0303311    64.89    0.0000    ***
  du_5         1.99335     0.0606622    32.86    8.24e-237 ***
  du_6         1.61537     0.0729014    22.16    8.67e-109 ***

   ⋮
```

```
Model 2: Fixed-effects, using 336 observations
Included 48 cross-sectional units
Time-series length = 7
Dependent variable: fatality
Robust (HAC) standard errors

            coefficient   std. error     z      p-value
  ------------------------------------------------------------
  const        2.37707      0.149797    15.87    1.04e-56 ***
  beertax     -0.655874     0.291856    -2.247   0.0246   **

    ⋮
```

时间固定效应回归

Time fixed effects regression

- 个体间相同但随时间变化的变量可以用时**间固定效应（time fixed effects）**控制。例如：新车安全性能的提高。

- 用 S_t 表示不可观测的随时间变化但个体间相同的效应。将 S_t 加入个体固定效应回归模型可得

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + \beta_3 S_t + u_{it}$$

- 当回归模型中仅有时间效应时，模型简化为

$$Y_{it} = \beta_1 X_{it} + \lambda_t + u_{it}$$

其中 $\lambda_1, \ldots, \lambda_T$ 为**时间固定效应**。和个体固定效应一样，这个模型可以用 $T-1$ 个二值变量表示，并用 OLS 估计。

个体和时间固定效应

- 同时加入个体和时间固定效应的回归模型可以写为

$$Y_{it} = \beta_1 X_{it} + \alpha_i + \lambda_t + u_{it}$$

其中 α_i 为个体固定效应，λ_t 为时间固定效应。

- 这个模型也可以用二值变量等价的表示为

$$\begin{aligned} Y_{it} =& \beta_0 + \beta_1 X_{it} + \gamma_2 D2_i + \cdots + \gamma_n Dn_i \\ &+ \delta_2 B2_t + \cdots + \delta_T BT_t + u_{it} \end{aligned}$$

其中共有 $n + T$ 个系数需要估计。

估计

- 加入个体和时间固定效应的回归模型有三种处理方法：
  1. 加入个体和时间的二值变量；
  2. 减去个体和时间平均值；
  3. 加入时间二值变量后，减去每个变量的个体平均值。
- 以上每一种处理后都可以用 OLS 进行估值。
- gretl 的 panel 命令采用的是第三种方法，具体为

```
panel fatality const beertax --time-dummies --robust
```

```
panel fatality const beertax --time-dummies --robust
```

```
Model 1: Fixed-effects, using 336 observations
Included 48 cross-sectional units
Time-series length = 7
Dependent variable: fatality
Robust (HAC) standard errors

             coefficient   std. error      z       p-value
  ------------------------------------------------------------
const         2.42847      0.201688      12.04     2.17e-33 ***
beertax      -0.639980     0.357078      -1.792    0.0731   *
dt_2         -0.0799029    0.0350861     -2.277    0.0228   **
dt_3         -0.0724206    0.0438809     -1.650    0.0989   *
dt_4         -0.123976     0.0460559     -2.692    0.0071   ***
dt_5         -0.0378645    0.0570604     -0.6636   0.5070
dt_6         -0.0509021    0.0636084     -0.8002   0.4236
dt_7         -0.0518038    0.0644023     -0.8044   0.4212
```

TABLE 10.1 Regression Analysis of the Effect of Drunk Driving Laws on Traffic Deaths

Dependent variable: traffic fatality rate (deaths per 10,000).

Regressor	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Beer tax	0.36 (0.05) [0.26, 0.46]	−0.66 (0.29) [−1.23, −0.09]	−0.64 (0.36) [−1.35, 0.07]	−0.45 (0.30) [−1.04, 0.14]	−0.69 (0.35) [−1.38, 0.00]	−0.46 (0.31) [−1.07, 0.15]	−0.93 (0.34) [−1.60, −0.26]
Drinking age 18		0.10		0.03 (0.07) [−0.11, 0.17]	−0.01 (0.08) [−0.17, 0.15]		0.04 (0.10) [−0.16, 0.24]
Drinking age 19				−0.02 (0.05) [−0.12, 0.08]	−0.08 (0.07) [−0.21, 0.06]		−0.07 (0.10) [−0.26, 0.13]
Drinking age 20				0.03 (0.05) [−0.07, 0.13]	−0.10 (0.06) [−0.21, 0.01]		−0.11 (0.13) [−0.36, 0.14]
Drinking age						0.00 (0.02) [−0.05, 0.04]	
Mandatory jail or community service?				0.04 (0.10) [−0.17, 0.25]	0.09 (0.11) [−0.14, 0.31]	0.04 (0.10) [−0.17, 0.25]	0.09 (0.16) [−0.24, 0.42]
Average vehicle miles per driver				0.008 (0.007)	0.017 (0.011)	0.009 (0.007)	0.124 (0.049)
Unemployment rate				−0.063 (0.013)		−0.063 (0.013)	−0.091 (0.021)
Real income per capita (logarithm)				1.82 (0.64)		1.79 (0.64)	1.00 (0.68)
Years	1982–88	1982–88	1982–88	1982–88	1982–88	1982–88	1982 & 1988 only
State effects?	no	yes	yes	yes	yes	yes	yes
Time effects?	no	no	yes	yes	yes	yes	yes
Clustered standard errors?	no	yes	yes	yes	yes	yes	yes

固定效应回归假设

个体固定效应回归模型

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}, \quad i = 1, \ldots, n,\ t = 1, \ldots, T$$

的假设条件为：

1. u_{it} 的条件均值为零：$E(u_{it} \mid X_{i1}, \ldots, X_{iT}, \alpha_i) = 0$ 。
2. $(X_{i1}, \ldots, X_{iT}, u_{i1}, \ldots, u_{iT}), i = 1, \ldots, n$ 是从其联合总体中抽取的 i.i.d. 样本。
3. 大异常值不太可能出现。
4. 不存在完全多重共线性。

当包含多个可观测回归变量时，X_{it} 应该替换为 $X_{1,it}, \ldots, X_{k,it}$ 。

异方差和自相关稳健标准误

Heteroskedasticity- and autocorrelation-consistent (HAC) standard errors

- 固定效应回归假设 2 成立时，不同个体对应的变量相互独立，但同一个体内的变量 X_{it} 可以夸时间相关。

- 若 X_{is} 和 X_{it} 相关，即给定个体的 X_{it} 跨时间相关，则称 X_{it} 为**自相关（autocorrelated）**或**序列相关（serially correlated）**。例如：啤酒税存在较强的自相关；误差项也有可能存在自相关。

- 存在自相关时，估计量依然为非偏，但用普通的异方差稳健方法计算的标准误会产生偏差。

- 同时适用于异方差和自相关的标准误被称为**异方差和自相关一致标准误（Heteroskedasticity- and autocorrelation-consistent (HAC) standard errors）**，本书中使用的为其中一种**群聚标准误（clustered standard errors）**。在 gretl 中设为 Arellano。

固定效应与随机效应

- 在固定效应模型中，我们暗中假定个体固定效应 α_i 为常数，因此可以当作参数进行估计。让这个估计可行的一个前提条件是回归变量 X_{it} 随时间变化而变化。如果 $X_{it} = X_i$，例如性别等不随时间变化而变化的变量，则无法估计个体固定效应。

- 另一种针对不可观测的个体效应的方法是**随机效应回归（random effects regression）**。随机效应回归模型假设不可观测变量 Z_i 为随机变量，因此可以将 α_i 假设为随机变量而非常数。同时，若 α_i 和回归变量 X_{it} 不相关，则可以令 $v_{it} = \alpha_i + u_{it}$ 为包含不可观测个体效应的误差项。此时回归模型可以表示为

$$Y_{it} = \beta_0 + \beta_1 X_{it} + v_{it}$$

  随机效应回归模型的估计需要用到**广义最小二乘法（generalized least squares, GLS）**。

- 在实际应用中大部分问题都更适合用固定效应回归进行分析，因此本书中没有涉及随机效应回归。感兴趣的同学可以参考 Dougherty (2016), 5th, Chapter 14。

课后练习（不需提交）

- 阅读 10.6 节并尝试在 gretl 中复制表 10.1 中的回归结果。
- 注意事项：
  - “饮酒年龄” 在数据中对应的变量是 `mlda` 。你需要自己将其转变为二值变量。
  - “强制监禁或者强制社区服务？” 在数据中对应的变量是 `jaild` 和 `comserd` 。你需要自己将这两个二值变量转变为一个二值变量。
  - “每个驾驶员平均行车里程” 在数据中对应的变量是 `vmiles` ，在使用前应当除以 1000 。
  - “失业率” 在数据中对应的变量是 `unrate` ，不是 `unus` 。

拓展阅读

1. Ruhm, C. (1996). Alcohol policies and highway vehicle fatalities, *Journal of Health Economics*, 15:435-454.

2. Dougherty, C. (2016). Chapter 14: Introduction to panel data models, in *Introduction to Econometrics*, Fifth edition, Oxford University Press.