多变量回归模型

OLS with multiple regressors

杨点溢

Department of Government
London School of Economics and Political Science

Spark 社科量化系列课程

本节课内容

单自变量回归模型 Linear Regression with One Regressor

$$Y_i = \alpha + \beta X_i + u_i$$

- 误差项 Error Term u_i 包含了其他影响 Y 但是没有包含在回归函数里的变量
- 这些变量叫做遗漏变量(Omitted Variables)
- 但是在回归模型中我们不能涵盖所有影响 Y 的变量 $\Rightarrow$ 所以这是个问题吗？
- 而且，须知我们需要 u 和 X 独立才能正确地将 β 解读为处理效应 (X 对 Y 的影响)

OVB /1

- 假设我们有一个遗漏变量 Z.
  - 或者一个包含很多遗漏变量的向量 Vector Z.
- 当有以下情况时，回归模型中遗漏 Z 将会是一个问题：
  - Z 被包含在 u 中 $(Z \ Y)$；
  - $corr(X, Z) \neq 0$，也就是说，Z 和被包含的变量（“处理变量” X）相关。
- 如果 1 和 2 都满足，那么 u 将与 X 相关 $\Rightarrow$ 选择偏差 (Selection Bias) 不是 0.
- 在回归分析中，这种偏差叫做遗漏变量偏差 (OVB)(Omitted Variable Bias)，是选择偏差的一种

OVB /2

- 假设真实的模型是：

$$Y_i = \alpha + \beta X_i + \gamma Z_i + u_i$$

注：u_i 独立于 X_i，即 $cov(u_i, X_i) = 0$

- 但是如果你估计的是：

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + \eta_i$$

- 你得到的将是：

$$\hat{\beta} = \frac{cov(Y_i, X_i)}{Var(Xi)}$$

OVB /3

- 我们再把“对”的模型带入，可得

$$\begin{aligned}\hat{\beta} &= \frac{cov(\alpha + \beta X_i + \gamma Z_i + u_i, X_i)}{var(X_i} \\ &= \frac{\beta var(X_i) + \gamma cov(X_i, Z_i) + cov(u_i, X_i)}{var(X_i)} \\ &= \beta + \gamma \frac{cov(X_i, Z_i)}{var(X_i)}\end{aligned}$$

- BIAS = (coefficient of the excluded variable) × (coefficient of a regression of the excluded variable on the included variable)
  - 偏差 =(被遗漏变量的系数)×(Z 和 X 回归的系数)

案例：班级大小和留学生

TABLE 6.1 Differences in Test Scores for California School Districts with Low and High Student–Teacher Ratios, by the Percentage of English Learners in the District

	Student–Teacher Ratio < 20		Student–Teacher Ratio ≥ 20		Difference in Test Scores, Low vs. High STR	
	Average Test Score	*n*	Average Test Score	*n*	Difference	*t*-statistic
All districts	657.4	238	650.0	182	7.4	4.04
Percentage of English learners						
< 1.9%	664.5	76	665.4	27	−0.9	−0.30
1.9–8.8%	665.2	64	661.8	44	3.3	1.13
8.8–23.0%	654.9	54	649.7	50	5.2	1.72
> 23.0%	636.7	44	634.8	61	1.9	0.68

本节课内容

多变量回归模型 Multiple Regression Model

假设我们有两个自变量，人口方程是：

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, i = 1, ..., n$$

- Y 是因变量 (DV)
- X_1 和 X_2 是自变量 (regressors)
- α 是人口截距 (intercept)
- β_1 是控制 X_2 不变的情况下 X_1 的 1 个单位变化对 Y 的影响
- β_2 是控制 X_1 不变的情况下 X_2 的 1 个单位变化对 Y 的影响
- u_i 是回归残差项（包含其他遗漏变量）

系数的解读

- 想象 X_i 变化了 ΔX_1，控制 X_2 不变
- 变化之前的回归线为：

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2$$

- 变化之后为：

$$Y + \Delta Y = \alpha + \beta_1 (X_1 + \Delta X_1) + \beta_2 X_2$$

- 两者之差：

$$\Delta Y = \beta_1 \Delta X_1$$

- 因此

$$\beta_1 = \frac{\Delta Y}{\Delta X_1} \text{控制 } X_2 \text{ 不变}$$

$$\beta_2 = \frac{\Delta Y}{\Delta X_2} \text{控制 } X_1 \text{ 不变}$$

$$\alpha = \text{当} X_1 = X_2 = 0 \text{时} Y \text{的预测值}$$

举例：班级大小

- 简单回归——考试分数和学校师生比 (Student-Teacher Ratio)

$$TestScore = 698.9 - 2.28 \times STR$$

- 现在我们把英语非母语的人占区域的比重 (PctEL) 加进模型

$$TestScore = 686 - 1.1 \times STR - 0.65 \times PctEL$$

- 师生比 (STR) 的系数会怎样变化了呢？为什么呢？
  - 提示：$corr(STR, PctEL) = 0.19$

R 的产出： fixest

```
OLS estimation, Dep. Var.: testscr
Observations: 420
Standard-errors: Heteroskedasticity-robust
             Estimate Std. Error   t value  Pr(>|t|)
(Intercept) 686.032249   8.728224  78.59930 < 2.2e-16 ***
str          -1.101296   0.432847  -2.54431  0.011309 *
PctEL        -0.649777   0.031032 -20.93909 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 14.4   Adj. R2: 0.42368
```

Stata 的产出

```
reg testscr str pctel, robust;

Regression with robust standard errors                Number of obs =     420
                                                      F(  2,   417) =  223.82
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.4264
                                                      Root MSE      =  14.464

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -1.101296   .4328472    -2.54   0.011*   -1.95213   -.2504616
       pctel |  -.6497768   .0310318   -20.94   0.000*   -.710775   -.5887786
       _cons |   686.0322   8.728224    78.60   0.000*   668.8754    703.189
------------------------------------------------------------------------------
```

拟合程度 Measures of fit /1

- 计算拟合程度的方法和只有一个自变量时是一样的
  - Standard Error of the Regression (SER): 衡量 Y_i 相对于回归线的离散程度

$$SER = \sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2}$$

  - R^2 代表 Y 的方差中有多少被所有自变量解释

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

拟合程度 Measures of fit /2

- 新概念：”Adjusted R^2” 或者 $\bar{R}^2$ 考虑到自变量的数量：自变量越多，R^2 永远增加，而 $\bar{R}^2$ 就修复了这个问题

$$\bar{R}^2 = 1 - (\frac{n-1}{n-k-1})\frac{SSR}{TSS}$$

  - k 是自变量的数量
  - 自变量变多时，$\bar{R}^2$ 不见得增加
  - $\bar{R}^2 < R^2$ （但是当 n 很大时两者区别不大）

回到 R 的回归表格

	(1)
(Intercept)	686.032
	(8.728)
str	−1.101
	(0.433)
PctEL	−0.650
	(0.031)
Num.Obs.	420
R2	0.426
R2 Adj.	0.424
AIC	3439.1
BIC	3451.2
RMSE	14.41
Std.Errors	Heteroskedasticity-robust

- 对比 R^2 和 $\bar{R}^2$
- $Root\ MSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2}$.
  - 当 n 足够大时，$Root\ MSE \simeq SER$

OLS 的假设

假设 1: $E(e|X) = 0$:e_i 关于 $x_{1i}, x_{2i}, ..., x_{ki}$ 的条件分部平均数为 0.
假设 2: $(x_{1i}, x_{2i}, ..., x_{ki}, y_i)$ 是独立且相同分布的 (i.i.d) 随机变量.
假设 3: 异常值 (outliers) 不多. 有限峰度 (finite kurtosis):

- $E(x_i^4) < \infty . E(y_i^4) < \infty .$

假设 4: 没有完美多重共线性 perfect multicollinearity

- 假设 1 和只有一个自变量时一样，只不过排除了 OVB
- 假设 2 和 3 没有区别
- 假设 4 是新的

本节课内容

多重共线性 Multicollinearity /1

- 当一个自变量是另一个自变量的完美函数时，会出现完美多重共线性 (perfect multicollinearity).
- 在这种情况下，R/Stata 会自动取消 (drop) 掉一个变量，如果你看到有一个变量被拿出，你应该考虑一下发生了什么
- 情况 1：你把同一个变量加了两次

```
> lm_multi <- feols(testscr~str+str, data=data, vcov="HC1")
> lm_multi
OLS estimation, Dep. Var.: testscr
Observations: 420
Standard-errors: Heteroskedasticity-robust
             Estimate Std. Error  t value   Pr(>|t|)
(Intercept) 698.93295  10.364360 67.43619  < 2.2e-16 ***
str          -2.27981   0.519489 -4.38856 1.4467e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 18.5   Adj. R2: 0.04897
```

多重共线性 Multicollinearity /2

- 情况 2:

$$
\begin{aligned}
TestScore_i &= \alpha + \beta_1 D_i + \beta_2 + u_i \\
\text{where } D_i &= 1 \text{ if class size} \leq 20 \text{ and 0 otherwise} \\
B_i &= 1 \text{ if class size} > 20 \text{ and 0 otherwise} \\
&\Rightarrow \; B_i = 1 - D_i \; \Rightarrow \; \textbf{完美多重共线性}
\end{aligned}
$$

- 这种情况被叫做 “虚拟变量陷阱” (dummy variable trap).

虚拟变量陷阱 Dummy Variable Trap

- 假设我们有好几个虚拟变量，他们互斥且是有限的 (mutually exclusive and exhaustive)：每个观察值都属于且只属于一个分类 (every observation falls in one category only)
  - 例：LSE 的学生可以被分为本科生、硕士生和博士生。假如我们因此设置三个虚拟变量：U_i, M_i, P_i.
- 如果我们把三个分类都设成虚拟变量，并在回归模型 (regression) 中保留了常数项/截距 (constant term/intercept)，我们将会得到完美多重共线性 (multicollinearity)
- 解决办法
  - 省略一个分类（比如说本科生）
    - 作为参考组 (reference)
  - 或者省略截距/常数项

不完美多重共线性 imperfect multicollinearity

- 发生于两个自变量 (regressors) 高度相关（但 $|corr| \neq 1$）
- 这意味着 OLS 估计的系数 (coefficients)将会不精确 (imprecisely estimated)
  - 说人话：SE 很大
  - X_1 的系数是控制 X_2 不变的情况下，X_1 的影响。如果 X_1 和 X_2 高度相关，那么控制住 X_2 不变时，X_1 也很难有变化空间 $\Rightarrow var(\hat{\beta}_1)$ 会很高 $\Rightarrow SE(\hat{\beta}_1)$ 会高
  - 数据很难向我们提供 X_1 变化但是 X_2 不变的信息

本节课内容

OLS 估计量的样本分布 (sampling distribution)

- 基本跟之前讲的一样:
  - $E(\hat{\beta}_1) = \beta_1$
  - $var(\hat{\beta}_1)$ 和 n 是反比例关系 (inversely proportional).
  - $\hat{\beta}_1 \xrightarrow{p} \beta_1$
  - $\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{var(\hat{\beta}_1)}} \sim N(0,1)$ when n is large.
- Same for $\hat{\beta}_2, ..., \hat{\beta}_k$

对于单个系数 (coefficient) 的检验

- 使用平常的 t 统计量
- 使用置信区间 (confidence interval): $\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)$
- 对于 $\hat{\beta}_2, ..., \hat{\beta}_k$ 同理
- 跟之前讲的完全一样

联合假说 (joint hypotheses) 检验

- 设 *Expn* =expenditure per pupil(每学生教育支出)，考虑以下模型：

$$TestScore = \alpha + \beta_1 STR + \beta_2 Expn + \beta_3 PctEL + u$$

联合假说 (joint hypotheses) 检验

- 设 *Expn* =expenditure per pupil(每学生教育支出)，考虑以下模型：

$$TestScore = \alpha + \beta_1 STR + \beta_2 Expn + \beta_3 PctEL + u$$

- 我们可能对如下假说感兴趣："学校资源不重要"以及"学校资源重要"。

联合假说 (joint hypotheses) 检验

- 设 *Expn* =expenditure per pupil(每学生教育支出)，考虑以下模型：

$$TestScore = \alpha + \beta_1 STR + \beta_2 Expn + \beta_3 PctEL + u$$

- 我们可能对如下假说感兴趣："学校资源不重要"以及"学校资源重要"。
- 普遍化：对于多个自变量系数 (coefficients)，我们可能对联合假说的检验感兴趣：

$$\begin{aligned} H_0 &: \beta_1 = 0 \text{ and } \beta_2 = 0 \\ \text{vs } H_1 &: \text{either } \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \text{ or both} \end{aligned}$$

联合假说 (joint hypotheses) 检验

- 设 *Expn* =expenditure per pupil(每学生教育支出)，考虑以下模型：

$$TestScore = \alpha + \beta_1 STR + \beta_2 Expn + \beta_3 PctEL + u$$

- 我们可能对如下假说感兴趣："学校资源不重要"以及"学校资源重要"。
- 普遍化：对于多个自变量系数 (coefficients)，我们可能对联合假说的检验感兴趣：

$$\begin{aligned} H_0 &: \beta_1 = 0 \text{ and } \beta_2 = 0 \\ \text{vs } H_1 &: \text{either } \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \text{ or both} \end{aligned}$$

- 联合假说 (joint hypotheses)明确了两个或以上系数的值。
  - 或者说，对两个或多个系数施加限制 (imposes a restriction on two or more coefficients)

联合假说 (joint hypotheses) 检验

- 设 *Expn* =expenditure per pupil(每学生教育支出)，考虑以下模型：

$$TestScore = \alpha + \beta_1 STR + \beta_2 Expn + \beta_3 PctEL + u$$

- 我们可能对如下假说感兴趣："学校资源不重要"以及"学校资源重要"。
- 普遍化：对于多个自变量系数 (coefficients)，我们可能对联合假说的检验感兴趣：

$$H_0 \ : \ \beta_1 = 0 \text{ and } \beta_2 = 0$$
$$\text{vs } H_1 \ : \ \text{either } \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \text{ or both}$$

- 联合假说 (joint hypotheses)明确了两个或以上系数的值。
  - 或者说，对两个或多个系数施加限制 (imposes a restriction on two or more coefficients)
  - 我们常用 q 来表示限制的数量 (number of restrictions)。在这个例子中，$q = 2$.

我们为什么不能一个一个地分别检验这些系数呢？

- 你可能非常想通过单个系数的 t 统计量来验证联合假说
  - 在任意 t 统计量大于 1.96 时拒绝 H_0.

我们为什么不能一个一个地分别检验这些系数呢？

- 你可能非常想通过单个系数的 t 统计量来验证联合假说
  - 在任意 t 统计量大于 1.96 时拒绝 H_0.
- 但是如果我们这么做的话，我们的显著性（假阳的概率）还会是 0.05 吗？

$$
\begin{aligned}
&Pr[|t_1| > 1.96 \text{ and/or } |t_2| > 1.96] = \\
&Pr[|t_1| > 1.96, |t_2| > 1.96] + Pr[|t_1| > 1.96, |t_2| \leq 1.96] \\
&\quad + Pr[|t_1| \leq 1.96, |t_2| > 1.96] \\
=&Pr[|t_1| > 1.96] \times Pr[|t_2| > 1.96] + Pr[|t_1| > 1.96] \times Pr[|t_2| \leq 1.96] \\
&\quad + Pr[|t_1| \leq 1.96] \times Pr[|t_2| > 1.96] \\
=&0.05 \times 0.05 + 0.05 \times 0.95 + 0.95 \times 0.05 \\
=&0.0975
\end{aligned}
$$

我们为什么不能一个一个地分别检验这些系数呢？

- 你可能非常想通过单个系数的 t 统计量来验证联合假说
  - 在任意 t 统计量大于 1.96 时拒绝 H_0.
- 但是如果我们这么做的话，我们的显著性（假阳的概率）还会是 0.05 吗？

$$
\begin{aligned}
&Pr[|t_1| > 1.96 \text{ and/or } |t_2| > 1.96] = \\
&Pr[|t_1| > 1.96, |t_2| > 1.96] + Pr[|t_1| > 1.96, |t_2| \leq 1.96] \\
&\quad + Pr[|t_1| \leq 1.96, |t_2| > 1.96] \\
=&Pr[|t_1| > 1.96] \times Pr[|t_2| > 1.96] + Pr[|t_1| > 1.96] \times Pr[|t_2| \leq 1.96] \\
&\quad + Pr[|t_1| \leq 1.96] \times Pr[|t_2| > 1.96] \\
=&0.05 \times 0.05 + 0.05 \times 0.95 + 0.95 \times 0.05 \\
=&0.0975
\end{aligned}
$$

- 这么做的话实际的显著水平会是 0.0975，即 9.75%，高于我们想要的 5%。

F 检验 (F-test)

- 我们刚刚使用“常理 (common sense) 得到的显著水平（假阳概率）并不是 5%；这种分别检验的方式实际的显著水平（假阳概率）高于我们想要的 5%⇒ 不够保守。

F 检验 (F-test)

- 我们刚刚使用“常理 (common sense) 得到的显著水平（假阳概率）并不是 5%；这种分别检验的方式实际的显著水平（假阳概率）高于我们想要的 5%⇒ 不够保守。
- 与此同时，我们忽略了 t_1 和 t_2 之间的相关性
  - ⇒ 当 β_1 和 β_2 相关时。t_1 和 t_2 也会相关。

F 检验 (F-test)

- 我们刚刚使用“常理 (common sense) 得到的显著水平（假阳概率）并不是 5%；这种分别检验的方式实际的显著水平（假阳概率）高于我们想要的 5%⇒ 不够保守。
- 与此同时，我们忽略了 t_1 和 t_2 之间的相关性
  - ⇒ 当 β_1 和 β_2 相关时。t_1 和 t_2 也会相关。
- 要检验联合假说，我们需要一个新的统计量：F 统计量。当只有两个自变量时，

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho} t_1 t_2}{1 - \hat{\rho}^2}\right)$$

  - ρ 为 t_1 和 t_2 之间相关性的估计值。
  - F 统计量修正了（考虑到了）t_1 和 t_2 之间的相关性

卡方分布 Chi-square distribution

- 考略一个特殊情况： t_1 独立于 $t_2 \Rightarrow \hat{\rho} \xrightarrow{p} = 0$
- 那么

$$F = \frac{1}{2}(\frac{t_1^2 + t_2^2 - 2\hat{\rho} t_1 t_2}{1 - \hat{\rho}^2}) \simeq \frac{1}{2}(t_1^2 + t_2^2)$$

- 在这种情况下，F 统计量的大样本分布即为两个独立分布 (indepedently distributed) 的标准正态分布(standard normal distribution, or $N(0,1)$) 的平方的平均
- q 个独立的 $N(0.1)$ 的平方之和的分布叫做卡方分布 (χ_q^2)
- 当 n 足够大时，F 的分布近似为 χ_q^2/q
  - 即为 $\chi_q^2/q \sim F_{q,\infty}$
  - 当 $n > 100$ 时,F 分布基本等同于 χ_q^2/q

R 的 F 检验应用

```
> linearHypothesis(lm1, c("str=0", "expn_stu=0"), test="F")
Linear hypothesis test

Hypothesis:
str = 0
expn_stu = 0

Model 1: restricted model
Model 2: testscr ~ str + expn_stu + PctEL

  Res.Df Df      F   Pr(>F)
1    418
2    416  2 5.4337 0.004682 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Stata 的 F 检验应用

```
reg testscr str expn_stu pctel, r;

Regression with robust standard errors                Number of obs =     420
                                                      F(  3,   416) =  147.20
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.4366
                                                      Root MSE      =  14.353

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -.2863992   .4820728    -0.59   0.553    -1.234001     .661203
    expn_stu |   .0038679   .0015807     2.45   0.015     .0007607    .0069751
       pctel |  -.6560227   .0317844   -20.64   0.000    -.7185008   -.5935446
       _cons |   649.5779   15.45834    42.02   0.000     619.1917    679.9641
------------------------------------------------------------------------------
```

NOTE

```
test str expn_stu;

 ( 1)  str = 0.0
 ( 2)  expn_stu = 0.0

F(  2,    416) =    5.43
Prob > F =    0.0047
```

(Assumption)

The test command follows the regression

There are q=2 restrictions being tested

The 5% critical value for q=2 is 3.00

Stata computes the p-value for you

多系数的单限制检验 Testing single restrictions on multiple coefficients

$$Y_i = \alpha + \beta X_i + \gamma Z_i + u_i, \ \ i = 1, ..., n$$

- 思考以下假说：

$$H_0 : \beta_1 = \beta_2 \text{ vs } H_1 : \beta_1 \neq \beta_2$$

多系数的单限制检验 Testing single restrictions on multiple coefficients

$$Y_i = \alpha + \beta X_i + \gamma Z_i + u_i, \ \ i = 1, ..., n$$

- 思考以下假说：

$$H_0 : \beta_1 = \beta_2 \text{ vs } H_1 : \beta_1 \neq \beta_2$$

- 零假说对多个系数赋予了一个限制 $(q = 1)$ The null imposes a single restriction $(q = 1)$ on multiple coefficients.

多系数的单限制检验 Testing single restrictions on multiple coefficients

$$Y_i = \alpha + \beta X_i + \gamma Z_i + u_i, \ \ i = 1, ..., n$$

- 思考以下假说：
$$H_0 : \beta_1 = \beta_2 \text{ vs } H_1 : \beta_1 \neq \beta_2$$
- 零假说对多个系数赋予了一个限制 $(q = 1)$ The null imposes a single restriction $(q = 1)$ on multiple coefficients.
- 两个方法：
  - 把模型重新整理 (rearrange) 为对于单一系数的单限制检验 (详见 Stock and Watson, 2014, p.276)

多系数的单限制检验 Testing single restrictions on multiple coefficients

$$Y_i = \alpha + \beta X_i + \gamma Z_i + u_i, \ \ i = 1, ..., n$$

- 思考以下假说：

$$H_0 : \beta_1 = \beta_2 \text{ vs } H_1 : \beta_1 \neq \beta_2$$

- 零假说对多个系数赋予了一个限制 $(q = 1)$ The null imposes a single restriction $(q = 1)$ on multiple coefficients.
- 两个方法：
  - 把模型重新整理 (rearrange) 为对于单一系数的单限制检验 (详见 Stock and Watson, 2014, p.276)
  - 直接使用软件：
    - Stata: test str=expn
    - R: linearHypothesis(lm1, "str=expn_stu", test="F")

本节课内容

回到考试分数数据库

- 我们想要得到无偏 (unbiased) 地估计班级大小对考试分数的影响（控制学生和学校的特征不变）。

回到考试分数数据库

- 我们想要得到无偏 (unbiased) 地估计班级大小对考试分数的影响（控制学生和学校的特征不变）。
- 以上目标需要我们思考加入什么变量、跑怎样的回归模型

回到考试分数数据库

- 我们想要得到无偏 (unbiased) 地估计班级大小对考试分数的影响 (控制学生和学校的特征不变)。
- 以上目标需要我们思考加入什么变量、跑怎样的回归模型
- 我们需要在电脑上写代码前就把这些问题考虑清楚 $\Rightarrow$ 提前思考模型参数

回到考试分数数据库

- 我们想要得到无偏 (unbiased) 地估计班级大小对考试分数的影响（控制学生和学校的特征不变）。
- 以上目标需要我们思考加入什么变量、跑怎样的回归模型
- 我们需要在电脑上写代码前就把这些问题考虑清楚 $\Rightarrow$ 提前思考模型参数
  - 先明确一个“基准”(baseline/benchmark)模型，再想一些涵盖其他不同变量的替代 (alternative) 的模型

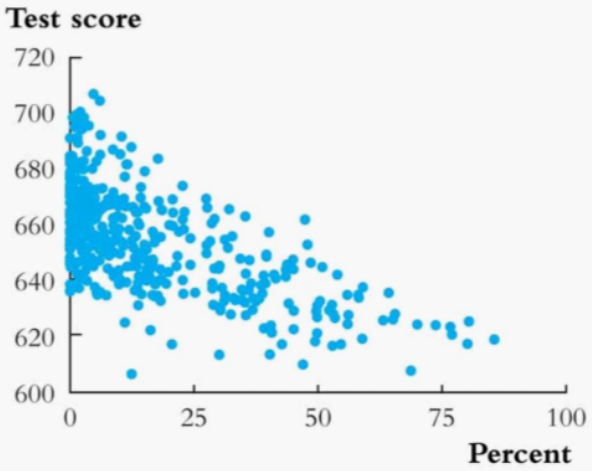
回到考试分数数据库

- 我们想要得到无偏 (unbiased) 地估计班级大小对考试分数的影响 (控制学生和学校的特征不变)。
- 以上目标需要我们思考加入什么变量、跑怎样的回归模型
- 我们需要在电脑上写代码前就把这些问题考虑清楚 $\Rightarrow$ 提前思考模型参数
  - 先明确一个“基准”(baseline/benchmark)模型，再想一些涵盖其他不同变量的替代 (alternative) 的模型
  - 加入一个新的变量会不会改变 $\hat{\beta}_1$？为什么？这个改变显著吗？

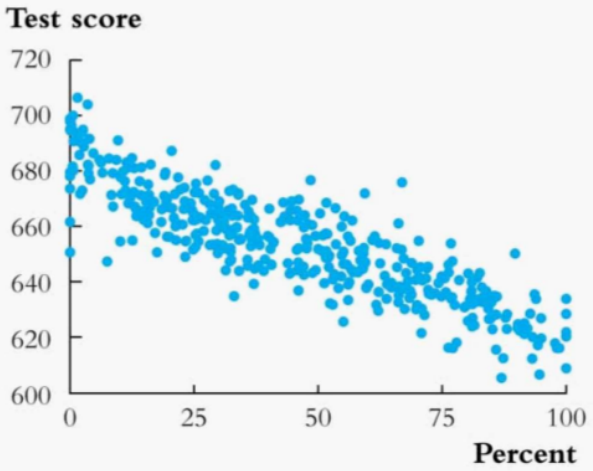
回到考试分数数据库

- 我们想要得到无偏 (unbiased) 地估计班级大小对考试分数的影响（控制学生和学校的特征不变）。
- 以上目标需要我们思考加入什么变量、跑怎样的回归模型
- 我们需要在电脑上写代码前就把这些问题考虑清楚 $\Rightarrow$ 提前思考模型参数
  - 先明确一个“基准”(baseline/benchmark)模型，再想一些涵盖其他不同变量的替代 (alternative) 的模型
  - 加入一个新的变量会不会改变 $\hat{\beta}_1$？为什么？这个改变显著吗？
  - 没有固定的公式：使用你的判断
  - 不要只想着最大化 R^2.

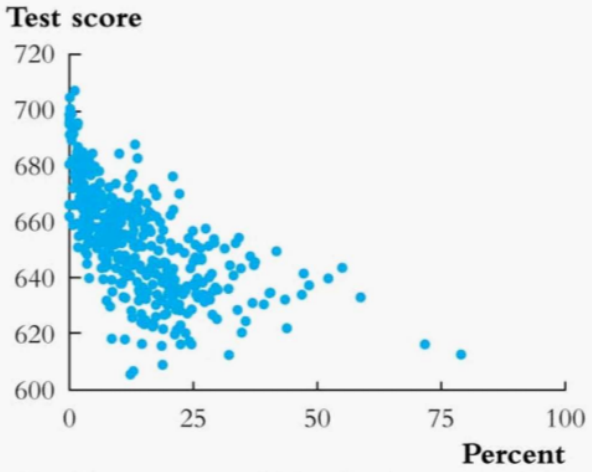
使用散点图



(a) Percentage of English language learners



(b) Percentage qualifying for reduced price lunch



(c) Percentage qualifying for income assistance

- 如果不明确，画上回归线和置信区间！

使用回归表格展示结果

- 我们会有好几个回归模型，我们想把他们的结果一并展示
- 国际惯例：做成个表格
- 一个回归结果表格应该包括：

使用回归表格展示结果

- 我们会有好几个回归模型，我们想把他们的结果一并展示
- 国际惯例：做成个表格
- 一个回归结果表格应该包括：
  - 估计的回归系数 (β)

使用回归表格展示结果

- 我们会有好几个回归模型，我们想把他们的结果一并展示
- 国际惯例：做成个表格
- 一个回归结果表格应该包括：
  - 估计的回归系数 (β)
  - 系数对应的标准误差 (se)
    - 有争议需不需要点星星
    - 有些模型没有 se，可以用置信区间代替

使用回归表格展示结果

- 我们会有好几个回归模型，我们想把他们的结果一并展示
- 国际惯例：做成个表格
- 一个回归结果表格应该包括：
  - 估计的回归系数 (β)
  - 系数对应的标准误差 (se)
    - 有争议需不需要点星星
    - 有些模型没有 se，可以用置信区间代替
  - 拟合程度 (measures of fit)
    - 通常是 R^2 或者 $\bar{R}^2$.

使用回归表格展示结果

- 我们会有好几个回归模型，我们想把他们的结果一并展示
- 国际惯例：做成个表格
- 一个回归结果表格应该包括：
  - 估计的回归系数 (β)
  - 系数对应的标准误差 (se)
    - 有争议需不需要点星星
    - 有些模型没有 se，可以用置信区间代替
  - 拟合程度 (measures of fit)
    - 通常是 R^2 或者 $\bar{R}^2$.
  - 观察量 (number of observations) (n)

使用回归表格展示结果

- 我们会有好几个回归模型，我们想把他们的结果一并展示
- 国际惯例：做成个表格
- 一个回归结果表格应该包括：
  - 估计的回归系数 (β)
  - 系数对应的标准误差 (se)
    - 有争议需不需要点星星
    - 有些模型没有 se，可以用置信区间代替
  - 拟合程度 (measures of fit)
    - 通常是 R^2 或者 $\bar{R}^2$.
  - 观察量 (number of observations) (n)
  - 如果需要，F 统计量
  - 其他重要的信息

学生分数结果表格 (Stock and Watson, 2014, p.287)

TABLE 7.1 Results of Regressions of Test Scores on the Student–Teacher Ratio and Student Characteristic Control Variables Using California Elementary School Districts

Dependent variable: average test score in the district.

Regressor	(1)	(2)	(3)	(4)	(5)
Student–teacher ratio (X_1)	−2.28** (0.52)	−1.10* (0.43)	−1.00** (0.27)	−1.31** (0.34)	−1.01** (0.27)
Percent English learners (X_2)		−0.650** (0.031)	−0.122** (0.033)	−0.488** (0.030)	−0.130** (0.036)
Percent eligible for subsidized lunch (X_3)			−0.547** (0.024)		−0.529** (0.038)
Percent on public income assistance (X_4)				−0.790** (0.068)	0.048 (0.059)
Intercept	698.9** (10.4)	686.0** (8.7)	700.2** (5.6)	698.0** (6.9)	700.4** (5.5)
Summary Statistics					
SER	18.58	14.46	9.08	11.65	9.08
$\overline{R}^2$	0.049	0.424	0.773	0.626	0.773
n	420	420	420	420	420

These regressions were estimated using the data on K-8 school districts in California, described in Appendix 4.1. Standard errors are given in parentheses under coefficients. The individual coefficient is statistically significant at the *5% level or **1% significance level using a two-sided test.

本节课内容

1. 遗漏变量偏差 Omitted Variable Bias(OVB)
2. 多变量回归模型 Multiple Regression Model
3. 多重共线性 Multicollinearity
4. 多变量 OLS 的假说检验 (hypothesis testing)
5. 案例：考试分数数据
6. **总结**

总结

- 多元线性回归可以分析在控制 X_2 不变的情况下，X_1 对 Y 的影响
- OVB 公式：

$$\begin{aligned} OVB = &(\text{coeff. of excluded variable}) \times \\ &(\text{coeff. of regression of excluded on included variable}) \end{aligned}$$

- 如果你可以测量到一个变量，你可以通过把他包含在你的模型里来减少选择偏差
- 但是我们往往不能观察/测量所有的变量
- 对于应该涵盖哪些变量，没有一个固定的配方——你得有自己的判断。

References I

Stock, J., & Watson, M. (2014). *Econometrics, Update PDF Ebook, Global Edtion*. Pearson Education, Limited. Retrieved July 19, 2023, from http://ebookcentral.proquest.com/lib/londonschoolecons/detail.action?docID=5174962