计量经济学

第七讲：多元线性回归

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

- 一元线性回归要点回顾
- 遗漏变量偏差
  - 什么是遗漏变量偏差
  - 遗漏变量偏差对 OLS 估计的影响
- 多元回归模型
  - 多元回归模型的 OLS 估计及假设
  - 多重共线性
- 假设检验与模型设定

一元线性回归要点回顾

线性回归模型

Linear regression model

- 一元线性回归模型

总体回归线 / 总体回归函数
population regression line/
population regression function

$$Y_i = \boxed{\beta_0 + \beta_1 X_i} + u_i$$

系数
coefficients

误差项
error term

因变量
dependent variable

自变量 / 回归变量
independent variable / regressor

OLS 估计量、预测值和残差

OLS estimator, predicted values, and residuals

- 斜率 β_1 和截距 β_0 的 OLS 估计量分别为

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

- OLS 预测值（predicted value）： $\hat{Y}_i = \boxed{\hat{\beta}_0 + \hat{\beta}_1 X_i}$
- 残差： $\hat{u}_i = Y_i - \hat{Y}_i$

样本回归线 / 样本回归函数
sample regression line/
sample regression function

最小二乘假设

The least square assumptions

一元回归模型

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \quad i = 1,2,\ldots,n$$

的最小二乘假设为：

1. 给定 X_i 时误差项 u_i 的条件分布均值为零：$E(u_i \mid X_i) = 0$
2. $(X_i, Y_i), i = 1,2,\ldots,n$ 为独立同分布
3. 不太可能出现大异常值：$0 < E(X_i^4) < \infty, 0 < E(Y_i^4) < \infty$

- 以上假设成立时，OLS 估计量是无偏估计量，若同时满足大样本条件，则 OLS 估计量具有一致性且其抽样分布为正态分布。

β_1 的假设检验

- β_1 的双边假设：

$$H_0 : \beta_1 = \beta_{1,0}; \quad H_1 : \beta_1 \neq \beta_{1,0}$$

- t 统计量为

$$t = \frac{\widehat{\beta}_1 - \beta_{1,0}}{\mathrm{SE}(\widehat{\beta}_1)},$$

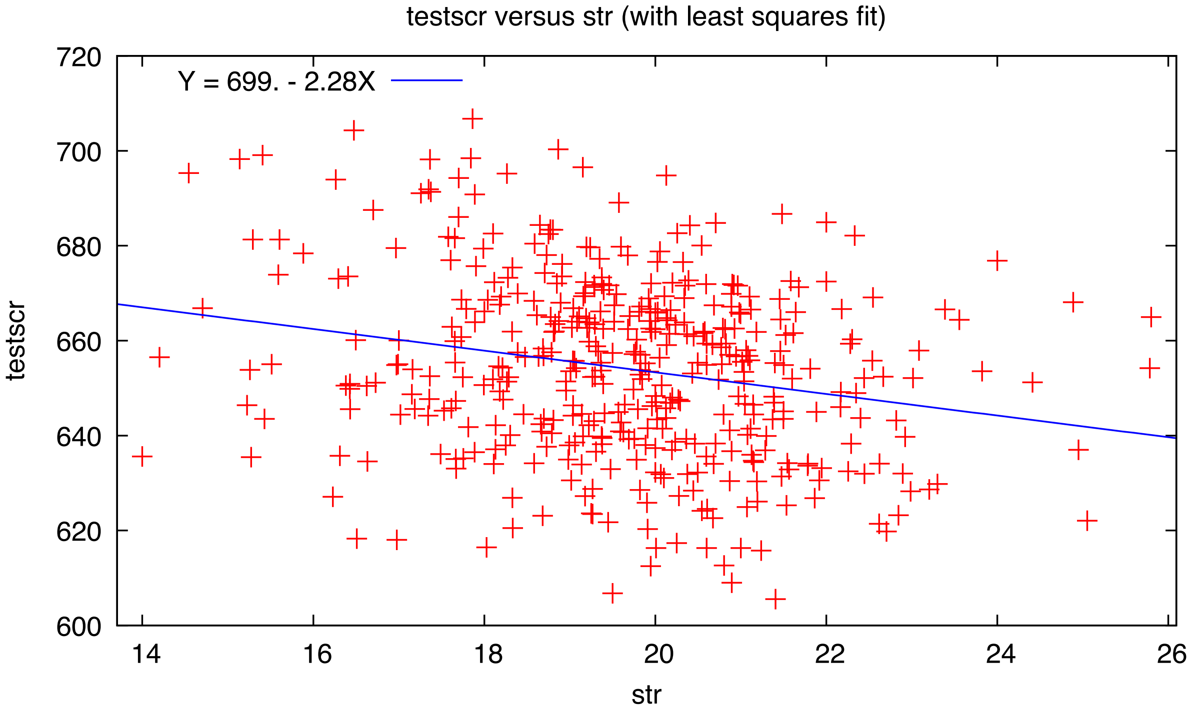
其中 $\mathrm{SE}(\widehat{\beta}_1) = \sqrt{\widehat{\sigma}^2_{\widehat{\beta}_1}}$, $\widehat{\sigma}^2_{\widehat{\beta}_1} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \widehat{u}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2}$

异方差稳健标准误（heteroskedasticity-robust SE, HC1）

遗漏变量偏差

STAR 数据集

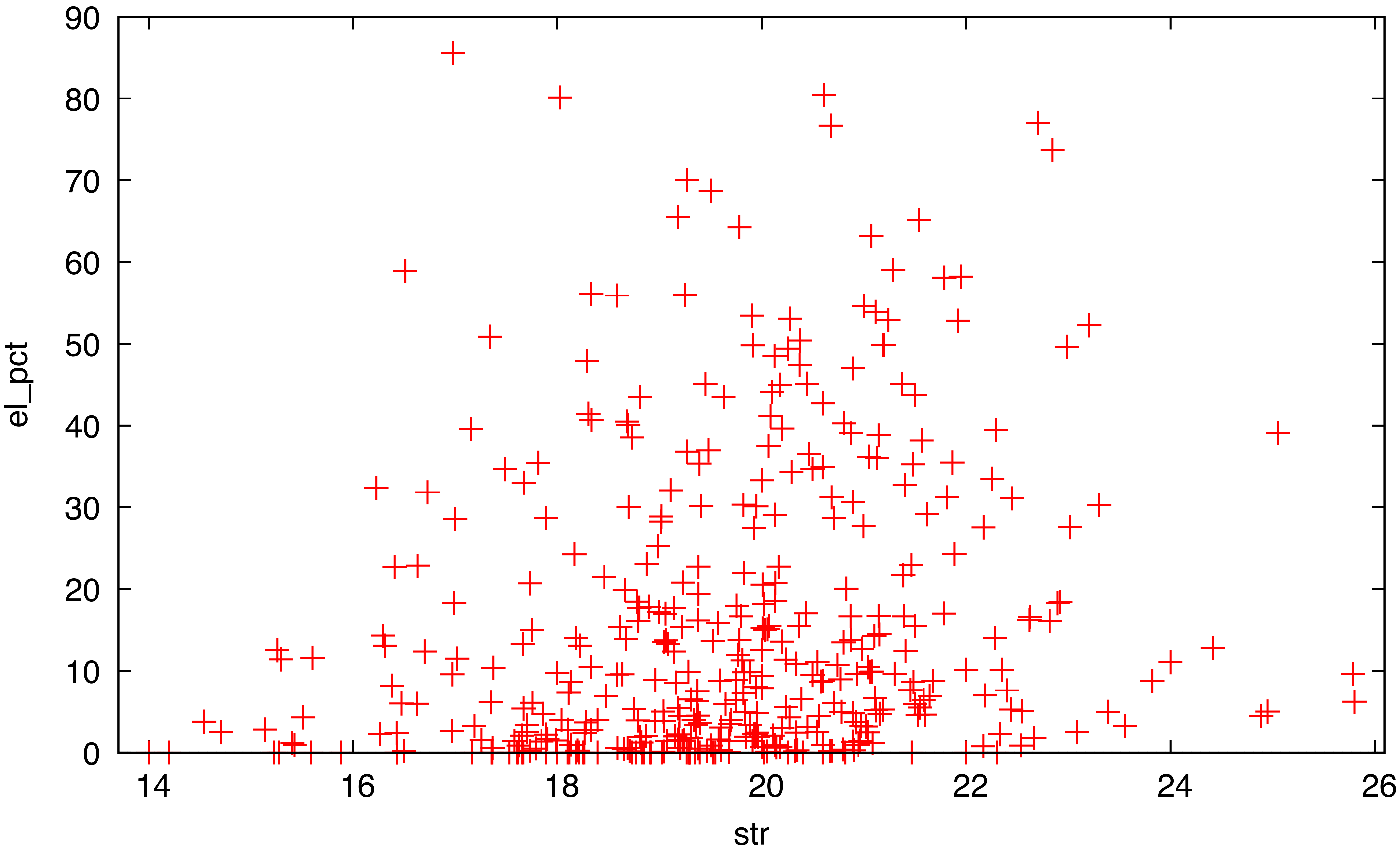
- STAR 数据集中可能影响 testscr 的变量
  - total enrollment
  - number of teachers
  - number of computers
  - computers per student
  - expenditures per student
  - **student teacher ratio**
  - percent of english learners
  - percent qualifying for reduced-price lunch
  - percent qualifying for CalWORKs
    (California Work Opportunities and Responsibility to Kids program)
  - district average income



学生教师比与仍在学习英语的学生比例

Student teacher ratio and percentage of English learners

两个变量间的相关系数为 0.188



当其他变量与 X 相关时

- 假设 u 中包含某个变量 Z，且 Z 和 X 正相关。同时假设 Z 的增加会引起 Y 的增加。

- 假设 X 对 Y 的影响为 $\beta_1 > 0$，即 X 增加一单位，Y 增加 β_1 单位。

- 当 X 增加 ΔX 时，Z 同时增加 ΔZ，此时 Y 的变化量为

$$\Delta Y = \beta_1 \Delta X + f(\Delta Z)$$

因此我们观察到的 Y 的变化比实际上由 X 引起的变化大，从而导致回归系数的估计值比真实值 β_1 大（overestimation）。

遗漏变量偏差

Omitted variable bias

- 如果回归变量与回归中漏掉的并对因变量起部分决定作用的某个变量（遗漏变量，omitted variable）相关，则 OLS 估计量有遗漏变量偏差（omitted variable bias）。

当满足下列条件时会产生遗漏变量偏差：

1. 遗漏变量和回归变量相关，
2. 遗漏变量是因变量的一个决定因素。

相关

$$Y_i = \beta_0 + \beta_1 X_i + \boxed{u_i}$$

其他变量

影响

例子：

1. 英语学习者百分比：满足 1 和 2；
2. 考试时间：不满足 1 ，但满足 2；
3. 每个学生的停车空间（教师停车场面积除以学生人数）：满足 1；但不满足 2。

遗漏变量偏差的影响

- 从第一条 OLS 假设 $E(u_i \mid X_i) = 0$ 可导出 $\mathrm{corr}(X_i, u_i) = 0$。
- 如果存在遗漏变量偏差，则

$$\mathrm{corr}(X_i, u_i) \neq 0 \quad \Rightarrow \quad E(u_i \mid X_i) \neq 0$$

- 若第一条 OLS 假设不成立，则 OLS 估计量是**有偏的**（因此称为遗漏变量偏差）。这个偏差即使在大样本下也不会消失，因此 OLS 估计量是**非一致的**。遗漏变量偏差下的 OLS 估计量满足

$$\widehat{\beta}_1 \xrightarrow{p} \beta_1 + \rho_{Xu}\frac{\sigma_u}{\sigma_X}, \quad \text{此处 } \rho_{Xu} = \mathrm{corr}(X_i, u_i)$$

多元回归模型

多元回归模型

Multiple regression model

- 多元线性回归模型（Linear regression model with multiple regressors）：

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_m X_{mi} + u_i$$

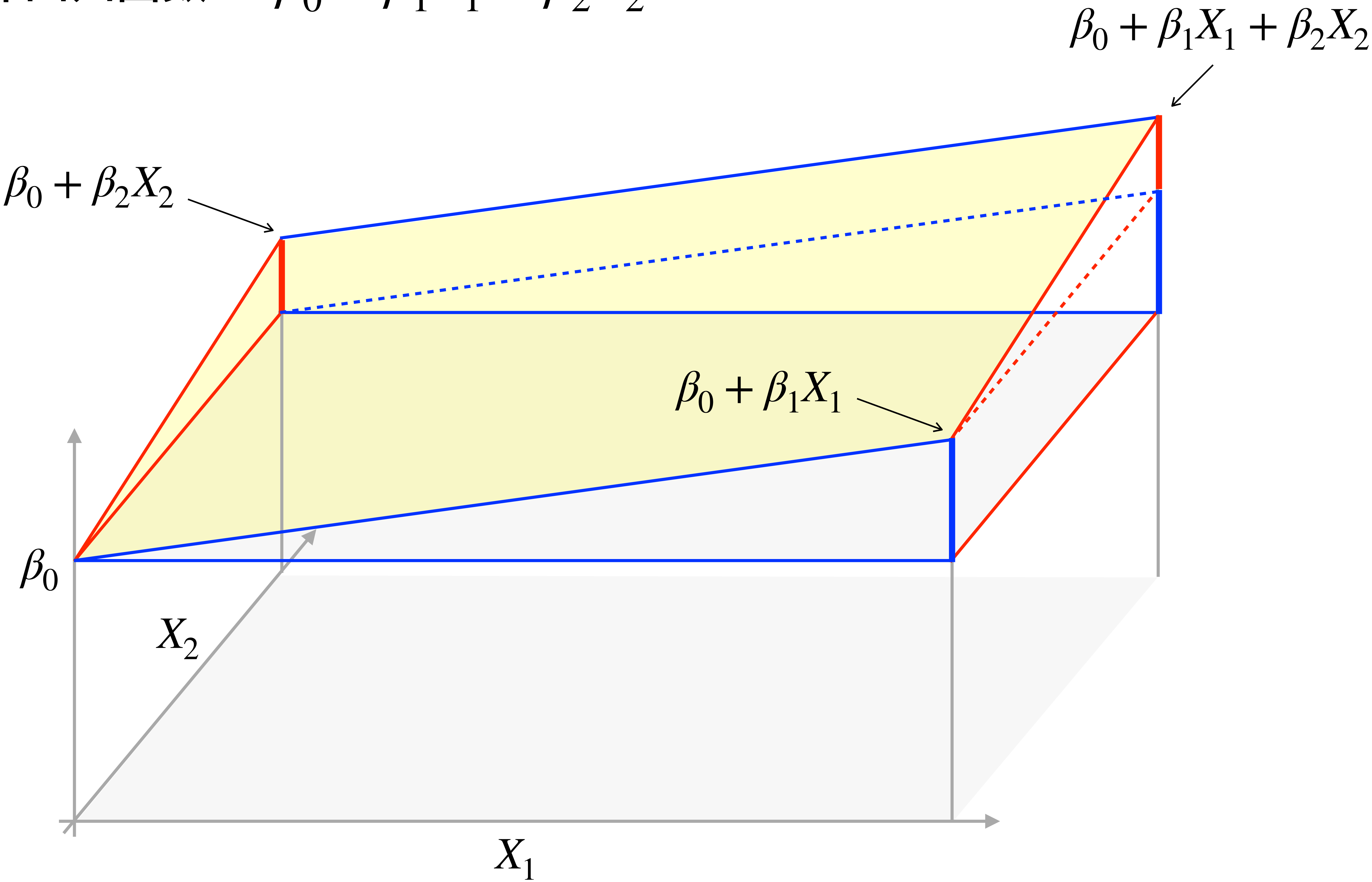
- 总体回归函数：

$$E(Y \mid X_1, X_2, \ldots, X_m) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m$$

- 截距 β_0： 当所有 X_k $(i = 1, \ldots, m)$ 都取零时 Y 的期望值。
- 系数 β_k： 保持其他 X 不变时， X_k 变化一个单位引起的 Y 的期望变化。
- 若 X_1 为主要关注的变量，则通常将其他自变量称为控制变量（control variable）。

总体回归函数的图形解释

总体回归函数 $= \beta_0 + \beta_1 X_1 + \beta_2 X_2$



OLS 估计量

- 回归系数的 OLS 估计量 $\widehat{\beta}_0, \widehat{\beta}_1, \ldots \widehat{\beta}_m$ 是令下列预测误差平方和最小的系数组合

$$\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{1i} - \ldots - b_m X_{mi})^2$$

- 求解方法：分别针对 b_k 求偏导，可得 $m+1$ 个一阶条件，即 $m+1$ 个线性方程。$\widehat{\beta}_0, \widehat{\beta}_1, \ldots \widehat{\beta}_m$ 满足该方程组。
- OLS 估计量的一般公式需要用矩阵表达，有兴趣的同学可以参考第十八章。

对 testscr 的多元回归分析

- 英语学习者百分比，`el_pct`，是另一个可以影响考试成绩的变量（testscr 是数学和阅读成绩的平均值，我们有理由猜测英语非母语学生的阅读成绩低于英语母语学生）。因此，回归模型可以扩展为：
  - 因变量：$Y = \texttt{testscr}$
  - 自变量：$X_1 = \texttt{str}, X_2 = \texttt{elpct}$
  - 回归模型：$\text{testscr}_i = \beta_0 + \beta_1 \text{str}_i + \beta_2 \text{elpct}_i + u_i$
- grerl 命令：

```
ols testscr const str el_pct --robust
```

多元回归结果

```
model1: OLS, using observations 1-420
Dependent variable: testscr
Heteroskedasticity-robust standard errors, variant HC1

             coefficient   std. error      z       p-value
  --------------------------------------------------------
  const       686.032      8.72822       78.60    0.0000    ***
  str          -1.10130    0.432847      -2.544   0.0109    **
  el_pct       -0.649777   0.0310318    -20.94    2.36e-97  ***

Mean dependent var   654.1565   S.D. dependent var   19.05335
Sum squared resid    87245.29   S.E. of regression   14.46448
R-squared            0.426431   Adjusted R-squared   0.423680
F(2, 417)            223.8229   P-value(F)           9.28e-67
Log-likelihood      -1716.561   Akaike criterion     3439.123
Schwarz criterion    3451.243   Hannan-Quinn         3443.913
```

一元回归结果

```
Model 1: OLS, using observations 1-420
Dependent variable: testscr
Heteroskedasticity-robust standard errors, variant HC1

             coefficient   std. error     z       p-value
  -------------------------------------------------------
  const       698.933      10.3644      67.44     0.0000    ***
  str          -2.27981     0.519489    -4.389    1.14e-05  ***
```

一元回归中的系数 β_1 估计值偏大（绝对值）

```
Mean dependent var   654.1565   S.D. dependent var   19.05335
Sum squared resid    144315.5   S.E. of regression   18.58097
R-squared            0.051240   Adjusted R-squared   0.048970
F(1, 418)            19.25943   P-value(F)           0.000014
Log-likelihood      -1822.250   Akaike criterion     3648.499
Schwarz criterion    3656.580   Hannan-Quinn         3651.693
```

R^2 和调整 R^2

- 在多元回归模型中，R^2 的定义依旧是

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = \frac{ESS}{TSS}$$
$$= 1 - \frac{\sum_{i=1}^{n}\hat{u}_i^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = 1 - \frac{SSR}{TSS}$$

除非增加的回归变量系数估值恰好为零，否则只要增加回归变量个数，SSR 就会增大，因此 R^2 就会增大。而实际上增加新的回归变量并不一定能提高模型的拟合程度，因此需要对 R^2 进行修正。

- 调整后的 R^2，即 $\overline{R}^2$ 为

$$\overline{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS}$$

k 为回归变量的个数

- $\overline{R}^2$ 的性质：
  - $\overline{R}^2$ 的取值永远小于 R^2
  - 回归变量增加时 $\overline{R}^2$ 不一定增加。$\overline{R}^2$ 的增减取决于新加入的回归变量对因变量影响的大小。

多元回归结果

```
model1: OLS, using observations 1-420
Dependent variable: testscr
Heteroskedasticity-robust standard errors, variant HC1

             coefficient   std. error      z       p-value
  --------------------------------------------------------
  const       686.032       8.72822       78.60    0.0000   ***
  str          -1.10130     0.432847      -2.544   0.0109   **
  el_pct       -0.649777    0.0310318    -20.94    2.36e-97 ***

Mean dependent var   654.1565   S.D. dependent var   19.05335
Sum squared resid    87245.29   S.E. of regression   14.46448
R-squared            0.426431   Adjusted R-squared   0.423680
F(2, 417)            223.8229   P-value(F)           9.28e-67
Log-likelihood      -1716.561   Akaike criterion     3439.123
Schwarz criterion    3451.243   Hannan-Quinn         3443.913
```

多元线性回归模型的 OLS 假设

多元回归模型

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_m X_{mi} + u_i, \quad i = 1,2,\ldots,n$$

的最小二乘假设为：

1. 给定 $X_{1i}, X_{2i}, \ldots, X_{mi}$ 时误差项 u_i 的条件分布均值为零：

$$E(u_i \mid X_{1i}, X_{2i}, \ldots, X_{mi}) = 0$$

（不存在遗漏变量）

2. $(X_{1i}, X_{2i}, \ldots, X_{mi}, Y_i),\ i = 1,2,\ldots,n$ 为独立同分布
3. 不太可能出现大异常值：

$$0 < E(X_{ki}^4) < \infty, k = 1,\ldots,m, \quad 0 < E(Y_i^4) < \infty$$

4. 不存在**完全多重共线性**（perfect multicollinearity） （保证能够计算 OLS 估计量）

- 以上假设成立时，OLS 估计量是无偏估计量，并在大样本条件下服从联合正态分布，且每个 $\hat{\beta}_k$ 服从 $N(\beta_j, \sigma^2_{\hat{\beta}_j})$， $k = 0,1,\ldots,m$。

多重共线性

Multicollinearity

- 多重共线性：当一个回归变量是其他回归变量的线性组合时，回归变量间是**完全多重共线的（perfect multicollinear）**，也称存在**完全多重共线性（perfect multicollinearity）**。
- 多重共线性的影响：无法计算 OLS 估计量。
- 多重共线性的例子：虚拟变量陷阱（dummy variable trap）
- 不完全（imperfect）多重共线性：一个回归变量和其他回归变量高度相关但不完全相关。
  - OLS 估计量保持非偏
  - OLS 估计量的标准误会增大，导致无法精确估计

多重共线性的检测

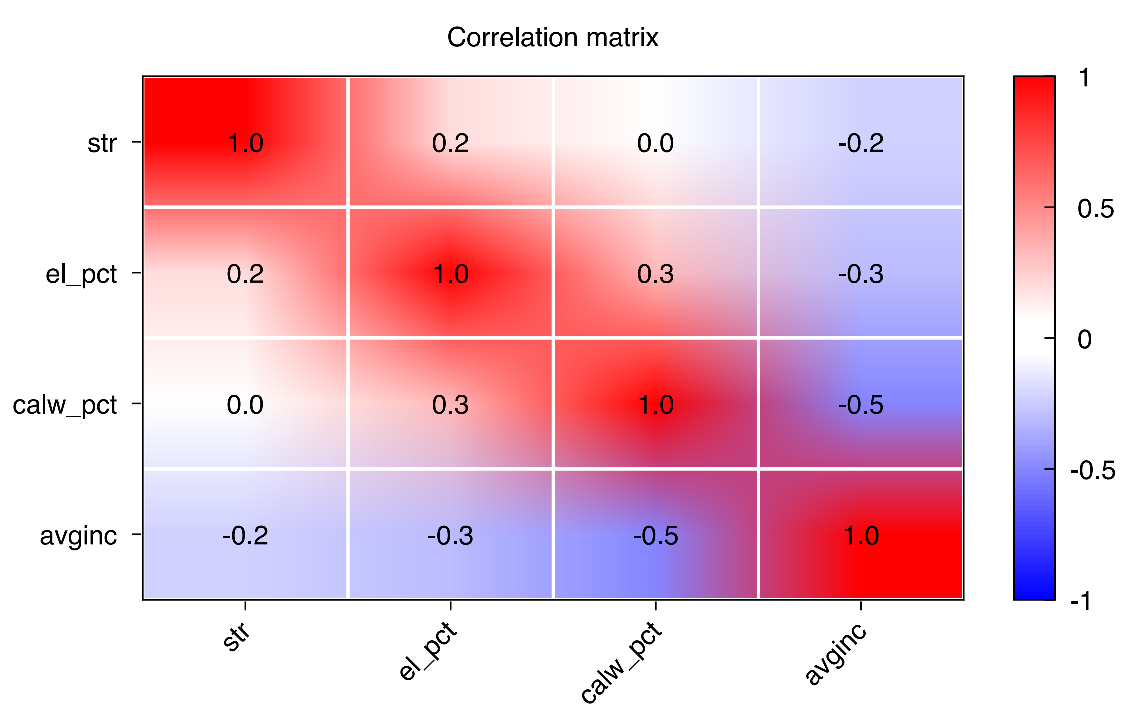
回归变量的样本相关矩阵

- 在 gretl 中可以用下列命令计算变量间的相关系数矩阵

```
corr str el_pct calw_pct avginc --plot=display
```

```
Correlation Coefficients, using the observations 1 - 420
5% critical value (two-tailed) = 0.0957 for n = 420

         str        el_pct      calw_pct        avginc
      1.0000        0.1876        0.0183       -0.2322   str
                    1.0000        0.3196       -0.3074   el_pct
                                  1.0000       -0.5127   calw_pct
                                                1.0000   avginc
```



多重共线性的检测

Variance Inflation Factors (VIF)

- Variance Inflation Factors (VIF)

$$\mathrm{VIF}_k = \frac{1}{1 - R_k^2},$$

此处 R_k^2 是用其他自变量回归 X_k 时得到的 R^2。

- 在运行 **`ols`** 命令后运行 **`vif`** 命令可得

```
Variance Inflation Factors
Minimum possible value = 1.0
Values > 10.0 may indicate a collinearity problem

         str    1.036
      el_pct    1.036

VIF(j) = 1/(1 - R(j)^2), where R(j) is the multiple correlation
coefficient between variable j and the other independent variables
```

存在多重共线性时该怎么办

- 存在完全多重共线性的回归模型无法估值，因此应该对模型进行调整，例如移除有问题的回归变量。

- 不完全多重共线性并不意味着回归模型存在问题。事实上回归系数间存在相关性是正常现象（见遗漏变量偏差的定义）。

- 回归系数间的相关性过大时会影响估值的精确程度，然而多大为 “过大” 也不存在唯一的标准。VIF 大于 10 的标准仅仅用于参考，不应该盲从。如果回归的主要目的是探索某个 X_k 对 Y 的因果效应，则其他回归变量的 VIF 可忽略。

- 增大样本容量也可以减小 OLS 估计量的标准误。

假设检验与模型设定

单个系数的假设检验

- β_k 的双边假设：

$$H_0 : \beta_k = \beta_{k,0}; \quad H_1 : \beta_k \neq \beta_{k,0}$$

- t 统计量为

$$t = \frac{\widehat{\beta}_k - \beta_{k,0}}{\mathrm{SE}(\widehat{\beta}_k)},$$

其中 $\mathrm{SE}(\widehat{\beta}_k)$ 可由计量软件算出，具体表达式见第十八章。

这里依然推荐使用异方差稳健标准误。

回归结果

```
model1: OLS, using observations 1-420
Dependent variable: testscr
Heteroskedasticity-robust standard errors, variant HC1

             coefficient   std. error      z       p-value
  --------------------------------------------------------
  const      686.032       8.72822        78.60    0.0000   ***
  str         -1.10130     0.432847       -2.544   0.0109   **
  el_pct      -0.649777    0.0310318     -20.94    2.36e-97 ***

Mean dependent var   654.1565   S.D. dependent var   19.05335
Sum squared resid    87245.29   S.E. of regression   14.46448
R-squared            0.426431   Adjusted R-squared   0.423680
F(2, 417)            223.8229   P-value(F)           9.28e-67
Log-likelihood      -1716.561   Akaike criterion     3439.123
Schwarz criterion    3451.243   Hannan-Quinn         3443.913
```

回归结果的表述

- 引入英语学习者百分比 `el_pct` 的回归模型为

$$\text{testscr}_i = \beta_0 + \beta_1 \text{str}_i + \beta_2 \text{elpct}_i + u_i$$

其回归结果可以写成

$$\widehat{\text{testscr}} = \underset{(8.7)}{686.0} - \underset{(0.43)}{1.10}\,\text{str} - \underset{(0.031)}{0.650}\,\text{elpct}$$

- 系数下方括号内为标准误。

联合假设检验

Joint hypothesis testing

- 对回归系数施加了两个或两个以上约束（restriction）的假设被称为联合假设（joint hypothesis）。

- 有 q 个约束的联合假设：

$$H_0: \beta_{k1} = \beta_{k1,0}, \beta_{k2} = \beta_{k2,0}, \ \ldots, \beta_{kq} = \beta_{kq,0}$$

$$H_1: H_0 \text{ 中 } q \text{ 个约束中的一个或多个约束不成立}$$

- 重要特例："总"回归的联合假设

$$H_0: \beta_1 = 0, \beta_2 = 0, \ \ldots, \beta_m = 0$$

$$H_1: \text{至少存在某个 } k \text{ 使 } \beta_k \neq 0, k = 1,2,\ldots,m$$

- 联合假设可以通过计算 F 统计量及其分布进行检验。q 个约束的异方差稳健 F 统计量在大样本下服从 $F_{q,\infty}$ 分布。

回归结果

```
model1: OLS, using observations 1-420
Dependent variable: testscr
Heteroskedasticity-robust standard errors, variant HC1

             coefficient   std. error      z      p-value
  --------------------------------------------------------
  const       686.032      8.72822       78.60    0.0000   ***
  str          -1.10130    0.432847      -2.544   0.0109   **
  el_pct       -0.649777   0.0310318    -20.94    2.36e-97 ***

Mean dependent var   654.1565   S.D. dependent var   19.05335
Sum squared resid    87245.29   S.E. of regression   14.46448
R-squared            0.426431   Adjusted R-squared   0.423680
F(2, 417)            223.8229   P-value(F)           9.28e-67
Log-likelihood      -1716.561   Akaike criterion     3439.123
Schwarz criterion    3451.243   Hannan-Quinn         3443.913
```

若选择异方差稳健标准误差，则 F 统计量也是异方差稳健的

涉及多个系数的单个约束检验

- 有时我们需要检验涉及两个或多个回归系数的单个约束。例如，如果理论指出 $\beta_1 = \beta_2$，则应针对下列假设进行检验：

$$H_0 : \beta_1 = \beta_2\ ; \quad H_1 : \beta_1 \neq \beta_2$$

- 因为只有一个约束条件，$q = 1$，F 统计量在大样本下服从 $F_{1,\infty}$ 分布。

- 在 gretl 中，我们可以先运行 OLS 回归，之后进行检验

```
ols testscr const str expn el_pct --robust

restrict
    b[2] - b[3] = 0
end restrict
```

回归结果

```
Model 2: OLS, using observations 1-420
Dependent variable: testscr
Heteroskedasticity-robust standard errors, variant HC1

             coefficient   std. error      z       p-value
  ---------------------------------------------------------
  const       649.578       15.4583      42.02     0.0000   ***
  str          -0.286399     0.482073    -0.5941   0.5524
  expn          3.86790      1.58072      2.447    0.0144   **
  el_pct       -0.656023     0.0317844  -20.64     1.21e-94 ***

Mean dependent var   654.1565   S.D. dependent var   19.05335
Sum squared resid    85699.71   S.E. of regression   14.35301
R-squared            0.436592   Adjusted R-squared   0.432529
F(3, 416)            147.2037   P-value(F)           5.20e-65
Log-likelihood      -1712.808   Akaike criterion     3433.615
Schwarz criterion    3449.776   Hannan-Quinn         3440.003

Excluding the constant, p-value was highest for variable 14 (str)
```

检验结果

```
Restriction:
 b[str] - b[expn] = 0

Test statistic: Robust F(1, 416) = 8.9403, with p-value =
0.00295511

Restricted estimates:

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const      685.822       11.3696       60.32    4.31e-208 ***
  str         -0.854052     0.459004     -1.861   0.0635    *
  expn        -0.854052     0.459004     -1.861   0.0635    *
  el_pct      -0.656690     0.0396393   -16.57    9.96e-48  ***

  Standard error of the regression = 14.5489
```

模型设定

Model specification

- 在应用研究中，我们经常需要决定将哪些变量纳入回归模型，即选择回归函数的具体形式（也包括后面将要涉及的函数型的选择）。
- 这是个非常具有挑战性的问题，且不存在适用于所有情况的规律。如果回归模型存在遗漏变量偏差，就要考虑遗漏变量的来源，而这取决于你对实证问题的了解程度。
  - 基于专业判断、经济理论和数据收集的方法选择核心或基础的回归变量集合，包括主要关注变量和控制变量。基于该回归变量基础集合的回归为**基础设定形式（base specification）**。
  - 同时列出候选的**备选设定形式（alternative specification）**，即备选的回归变量集合，包括其他无法通过理论判断是否应该加入模型的控制变量。
  - 如果主要关注的系数估计值在基础设定形式和所有备选设定形式中大小差不多，就有理由认为基础设定形式提供的估计值是可靠的，反之则表示基础设定形式可能存在遗漏变量偏差。
- 不要根据 R^2 或调整 R^2 的取值判断模型的设定是否合适。

控制变量的作用

- 在 OLS 假设下，所有的变量都具有相同的地位。哪个变量是主要变量，哪些是控制变量，完全取决于研究者的兴趣所在，在理论上没有区别。

- 如果将 OLS 假设 1 变更为下列**条件均值独立**假设，

$$E(u_i \mid X_{1i}, X_{2i}, \ldots, X_{mi}) = E(u_i \mid X_{2i}, \ldots, X_{mi})$$

  即 u_i 的条件均值不依赖于 X_{1i}，则 X_{1i} 为主要关注变量，其他自变量为控制变量。这时候，X_{1i} 的系数可以解释为因果关系，而其他系数则不能。

- 在条件均值独立假设下，控制变量依然可以和误差项相关，也就是说其系数可能受遗漏变量偏差影响。如果我们的研究目的不在控制变量的因果效应上，这种偏差也是可以接受的，因为他不影响主要结论。

用列表展示回归结果

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) [−3.30, −1.26]	−1.10 (0.43) [−1.95, −0.25]	−1.00 (0.27) [−1.53, −0.47]	−1.31 (0.34) [−1.97, −0.64]	−1.01 (0.27) [−1.54, −0.49]
Control variables					
Percentage English learners (X_2)		−0.650 (0.031)	−0.122 (0.033)	−0.488 (0.030)	−0.130 (0.036)
Percentage eligible for subsidized lunch (X_3)			−0.547 (0.024)		−0.529 (0.038)
Percentage qualifying for 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. Heteroskedasticity-robust standard errors are given in parentheses under coefficients. For the variable of interest, the student–teacher ratio, the 95% confidence interval is given in brackets below the standard error.

Table 4

Individual Contribution to the Public Good

Dep. var.: Individual contribution to the PGG				
	Model 1	Model 2	Model 3	Model 4
Northern Italy	1.213*	1.161**	1.066**	
	(0.580)	(0.432)	(0.429)	
Latitude				0.195***
				(0.057)
Individual choices over lotteries				
Strongly risk averse			0.806	0.806
			(0.572)	(0.569)
Risk neutral/risk loving			−0.921*	−0.895*
			(0.445)	(0.450)
Task comprehension (1 = low)		0.757	0.819	0.822
		(0.739)	(0.731)	(0.725)
Socio-demographic characteristics (control variables)	No	Yes	Yes	Yes
No. obs. (individuals)	372	372	372	372
R^2	0.015	0.085	0.101	0.106

Notes. OLS regression with standard errors robust for clustering at the session level (in parentheses). The dependent variable is the contribution of one participant averaged over all rounds of the PGG. The default category for risk preference is: moderately risk averse. Socio-demographic characteristics are listed in the main text. ***, **, and * indicate significance at the 1%, 5% and 10% level, respectively.

From Bigoni et al. (2016), *The Economic Journal*, 126:1318-1341

A guide for how to format tables and figures:
http://abacus.bates.edu/~ganderso/biology/resources/writing/HTWtablefigs.html

课后练习（不需提交）

- 阅读第 7.6 节并尝试复制其中的分析结果。
- 在 gretl 中，**modeltab** 命令可以帮助我们制作和书中类似的表格。具体用法参考 command reference。
  （注：gretl 无法替代你制作表格，最终表格的格式还需要手动编辑。）

  例：

```
modeltab free
ols testscr const str --robust --quiet
modeltab add
ols testscr const str el_pct --robust --quiet
modeltab add
modeltab show
```