线性回归中的假设检验

慧航

2025年8月

同方差下的标准误

- 对于线性回归模型

$$y_i = x_i'\beta + u_i$$

假设u_i是同方差的，那么

$$\sqrt{N}\left(\hat{\beta} - \beta\right) \overset{a}{\sim} \mathcal{N}\left(0, \sigma^2 \left[\mathbb{E}\left(x_i x_i'\right)\right]^{-1}\right)$$

- 渐进方差的估计：

$$\widehat{\mathbb{V}\left(\hat{\beta}\right)} = \frac{1}{N} s^2 \left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i'\right)^{-1} = s^2 \left(X'X\right)^{-1}$$

标准误：

$$\text{s.e.}\left(\hat{\beta}_k\right) = \sqrt{s^2 \left(X'X\right)^{-1}_{kk}}$$

异方差下的标准误

- 渐进分布

$$\sqrt{N}\left(\hat{\beta}-\beta\right) \overset{a}{\sim} \mathcal{N}\left(0, \left[\mathbb{E}\left(x_i x_i'\right)\right]^{-1} \mathbb{E}\left(u_i^2 x_i x_i'\right) \left[\mathbb{E}\left(x_i x_i'\right)\right]^{-1}\right)$$

- 渐进方差的估计量为：

$$\begin{aligned}
\widehat{\mathbb{V}\left(\hat{\beta}\right)} &= \frac{1}{N}\left[\frac{1}{N}\sum_{i=1}^{N}\left(x_i x_i'\right)\right]^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\hat{u}_i^2 x_i x_i'\right)\left[\frac{1}{N}\sum_{i=1}^{N}\left(x_i x_i'\right)\right]^{-1} \\
&= \left[\sum_{i=1}^{N}\left(x_i x_i'\right)\right]^{-1}\left(\sum_{i=1}^{N}\hat{u}_i^2 x_i x_i'\right)\left[\sum_{i=1}^{N}\left(x_i x_i'\right)\right]^{-1} \\
&= \left(X'X\right)^{-1}\left(\sum_{i=1}^{N}\hat{u}_i^2 x_i x_i'\right)\left(X'X\right)^{-1}
\end{aligned}$$

怀特异方差稳健标准误（White’s heteroscedasticity robust standard error）。

聚类标准误

很多时候我们使用的是分组数据：

$$y_{ig} = x'_{ig}\beta + u_{ig}$$

然而有时，误差项之间可能存在相关性：

$$\mathbb{C}\left(u_{ig}, u_{jg'}\right) \neq 0$$

可能的原因：组内某些不能观察到的变量对y_{ig}有影响。

- 例如，如果y_{ig}是班级内学生的成绩，那么g是班级，教师质量等都可能影响成绩。

很多时候不能单纯使用group dummy来解决

- 例如，学生努力程度的peer effect

此时，仅仅做异方差文件的s.e.是不够的。

聚类标准误

- 然而，有N个误差项，从而误差项之间的协方差矩阵为$N \times N$的矩阵，从而有$\frac{N(N-1)}{2}$个未知参数，无法一一估计
- 一个简单的假设：在组内，u_{ig}之间是相关的，然而不属于同一组的误差项之间不相关

$$\mathbb{E}\left(u_{ig}u_{jg'}\right) = \begin{cases} \sigma_{(ij)g} & g = g' \\ 0 & g \neq g' \end{cases}$$

聚类标准误

$$
\begin{bmatrix}
\sigma^2_{(11)1} & \cdots & \sigma^2_{(1N_1)1} & 0 & \cdots & 0 & 0 & \cdots & \cdots & 0 \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & & & \\
\sigma^2_{(N_1 1)1} & \cdots & \sigma^2_{(N_1 N_1)1} & 0 & \cdots & 0 & 0 & \cdots & \cdots & 0 \\
0 & \cdots & 0 & \sigma^2_{(11)2} & \cdots\cdots & \sigma^2_{(1N_2)2} & 0 & \cdots & \cdots & 0 \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & & & & \\
0 & \cdots & 0 & \sigma^2_{(N_2 1)2} & \cdots & \sigma^2_{(N_2 N_2)2} & 0 & \cdots & \cdots & 0 \\
 & & & & & & \ddots & & & \\
0 & 0 & 0 & 0 & 0 & 0 & \ddots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & 0 & \sigma^2_{(11)G} & \cdots\cdots & \sigma^2_{(1N_G)G} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & 0 & \vdots & \ddots & \vdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma^2_{(N_G 1)2} & \cdots & \sigma^2_{(N_G N_G)2}
\end{bmatrix}
$$

聚类标准误

- 在此设定下，可以使用聚类标准误：

$$\mathbb{V}\left(\hat{\beta}\right) = [X'X]^{-1}\left[\sum_g x'_g \hat{u}_g \hat{u}'_g x_g\right][X'X]^{-1}$$

- 如此，中间一项可以写为

$$\begin{aligned}
\sum_g x'_g \hat{u}_g \hat{u}'_g x_g &= \sum_g \left(\begin{bmatrix} x_{1g} & \cdots & x_{N_g g} \end{bmatrix} \begin{bmatrix} \hat{u}_{1g}^2 & \cdots & \hat{u}_{1g}\hat{u}_{N_g g} \\ \vdots & \ddots & \vdots \\ \hat{u}_{N_g g}\hat{u}_{1g} & \cdots & \hat{u}_{N_g g}^2 \end{bmatrix} \begin{bmatrix} x'_{1g} \\ \vdots \\ x'_{N_g g} \end{bmatrix} \right) \\
&= \sum_g \sum_{i\in g} \sum_{j\in g} \hat{u}_{ig}\hat{u}_{jg} x_{ig} x'_{jg} \\
&= \underbrace{\sum_g \sum_{i\in g} \hat{u}_{ig}^2 x_{ig} x'_{ig}}_{\text{异方差稳健}} + \underbrace{\sum_g \sum_{i\in g} \sum_{j\neq i} \hat{u}_{ig}\hat{u}_{jg} x_{ig} x'_{jg}}_{\text{组内协方差}}
\end{aligned}$$

聚类标准误

不止一组的情况：

- 如果是嵌套的，比如学校、班级，应该cluster在更高的一级
  - 至少要控制在与解释变量相同的层级。
- 如果不是嵌套的：
  - 控制一组的固定效应，cluster另外一组
  - two-way cluster，Cameron, Gelbach and Miller, 2006
- 现实中：
  - reg中的cluster()只能cluster一组
  - reghdfe可以cluster两组

聚类标准误

一般来说，cluster的标准误更大，因而参数的置信区间更大，越容易不显著。
cluster标准误比$i.i.d$的标准误扩大了：

$$\sqrt{1+\rho_x\rho_u\left(\bar{N}-1\right)}$$

其中ρ_x为解释变量的组内相关性，ρ_u为误差项组内相关性，而$\bar{N}$为平均的组大小。

固定效应和聚类标准误：一个实例

媒体对于俄罗斯大选的影响

Enikolopov, Petrova和Zhuravskaya(2011)研究了媒体（特别是独立的NTV电视台）对于俄罗斯大选的影响，他们的设定如下：

$$\text{vote}_{s,1999} = \beta_0 + \beta_1 \text{NTV}_{s,1999} + x'\beta + \delta_r + \epsilon_s$$

其中$\text{vote}_{s,1999}$为选区s在1999年选举中的结果，我们这里挑选对Unity党的投票结果；$\text{NTV}_{s,1999}$为估计的电视台覆盖率，δ_r为地区固定效应。
（fixed_effects_ntv.do）

小样本正态同方差条件下的分布

- 在**小样本**的条件下，我们使用了同方差假设以及正态性假设，并得到：

$$\frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2 \left(X'X\right)^{-1}_{kk}}} \sim \mathcal{N}\left(0,1\right)$$

- 在上式中，σ^2是未知的，我们可以使用其无偏估计s^2来代替，然而，由于使用s^2代替了σ^2，以上分布不再是标准正态分布，而是变成了自由度为$N-K$的t分布：

$$\frac{\hat{\beta}_k - \beta_k}{\text{s.e.}\left(\hat{\beta}_k\right)} = \frac{\hat{\beta}_k - \beta_k}{\sqrt{s^2 \left(X'X\right)^{-1}_{kk}}} \sim \text{t}\left(N-K\right)$$

正态同方差下的检验

- 基于以上结论，对于原假设：

$$H_0 : \beta_k = b$$

我们可以使用统计量

$$t_k = \frac{\hat{\beta}_k - b}{\text{s.e.}\left(\hat{\beta}_k\right)} \sim \text{t}\left(N - K\right)$$

对单个系数进行假设检验。

- 实践中，最经常遇到的情况是判断x_k对于y是否有影响，即原假设为：$H_0 : \beta_k = 0$，此时检验统计量为：

$$\frac{\hat{\beta}_k}{\text{s.e.}\left(\hat{\beta}_k\right)} \sim \text{t}\left(N - K\right)$$

即直接使用参数估计值除以标准误，并与$\text{t}\left(N - K\right)$的临界值进行比较就可以对原假设进行检验。

正态同方差下的检验

- 实践中，最经常遇到的情况是判断x_k对于y是否有影响，即原假设为：$H_0 : \beta_k = 0$，此时检验统计量为：

$$\frac{\hat{\beta}_k}{\text{s.e.}\left(\hat{\beta}_k\right)} \sim \text{t}\left(N - K\right)$$

  即直接使用参数估计值除以标准误，并与$\text{t}\left(N - K\right)$的临界值进行比较就可以对原假设进行检验。
- 在统计软件中，通常都会在计算$\hat{\beta}$之后，汇报标准误，并给出在原假设：$H_0 : \beta_k = 0$条件下的$p - value$。
- 在汇报回归结果时，通常将标准误s.e. $\left(\hat{\beta}_k\right)$标记在系数下面的括号中，同时使用“星星”颗数表示显著性水平

回归表格

VARIABLES	(1) y	(2) y	(3) y	(4) y
exer	-1.856***	2.912***	3.187***	3.059***
	(0.285)	(0.105)	(0.0990)	(0.0720)
gender				-10.05***
				(0.0720)
Constant	77.57***	70.02***	79.92***	79.96***
	(0.205)	(0.0915)	(0.0521)	(0.0483)
Observations	1,000	491	509	1,000
R-squared	0.041	0.613	0.671	0.953

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

大样本同方差条件下的分布

- 小样本下的检验仍然依赖于误差项的正态性假设，大样本条件下，可以使用CLT从而可以放弃误差项正态的假设
- 在大样本条件下，我们可以放弃正态性假设，有：

$$\sqrt{N}\left(\hat{\beta}-\beta\right) \overset{a}{\sim} \mathcal{N}\left(0, \sigma^2\left[\mathbb{E}\left(x_i x_i'\right)\right]^{-1}\right)$$

从而：

$$\sqrt{N}\left(\hat{\beta}_k-\beta_k\right) \overset{a}{\sim} \mathcal{N}\left(0, \sigma^2\left[\mathbb{E}\left(x_i x_i'\right)\right]_{kk}^{-1}\right)$$

大样本同方差条件下的分布

- 此外，由于：

$$s^2 \xrightarrow{p} \sigma^2, \frac{1}{N}\sum_{i=1}^{N} x_i x_i' \xrightarrow{p} \mathbb{E}\left(x_i x_i'\right)$$

从而：

$$s^2\left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i'\right)^{-1} = Ns^2\left(X'X\right)^{-1} \xrightarrow{p} \sigma^2\left[\mathbb{E}\left(x_i x_i'\right)\right]_{kk}^{-1}$$

- 根据Slutsky定理：

$$\frac{\hat{\beta}_k - \beta_k}{\text{s.e.}\left(\hat{\beta}_k\right)} = \frac{\sqrt{N}\left(\hat{\beta}_k - \beta_k\right)}{\sqrt{Ns^2\left(X'X\right)_{kk}^{-1}}} \overset{a}{\sim} \mathcal{N}(0,1)$$

大样本同方差条件下的检验

- 因而对于原假设$H_0:\beta_k=b$，我们可以使用统计量

$$\frac{\hat{\beta}_k-b}{\text{s.e.}\left(\hat{\beta}_k\right)}\overset{a}{\sim}\mathcal{N}(0,1)$$

  对原假设进行假设检验。
- 注意到以上的检验统计量与同方差、正态总体条件下的检验统计量是一模一样的，除了其分布变为标准正态分布。
- 实际上：
  - 如果K固定，当样本量足够大时，$\mathrm{t}(N-K)\xrightarrow{D}\mathcal{N}(0,1)$，此时使用$t$分布或者正态分布并无差异；
  - 另一方面，t分布的比标准正态分布具有更厚的尾巴，意味着在相同的显著性水平下，$\mathrm{t}(N-K)$的临界值总是大于标准正态分布的临界值，而在小样本时，t分布和正态分布都不精确，但是使用t分布却有着更小的犯第I类错误的概率。
  - 基于以上原因，在实践中，无论样本量大小，我们通常都使用$\mathrm{t}(N-K)$分布取临界值。

异方差稳健标准误

- 如果存在异方差，以上计算的标准误都是错误的
- 怀特异方差稳健标准误：

$$\widehat{\mathbb{V}\left(\hat{\beta}\right)} = \left(X'X\right)^{-1}\left(\sum_{i=1}^{N}\hat{u}_i^2 x_i x_i'\right)\left(X'X\right)^{-1}$$

- 在异方差的条件下，如果我们使用异方差稳健标准误，那么：

$$\frac{\hat{\beta}_k - \beta_k}{\text{s.e.}\left(\hat{\beta}_k\right)} = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\widehat{\mathbb{V}\left(\hat{\beta}\right)}_{kk}}} \overset{a}{\sim} \mathcal{N}\left(0,1\right)$$

仍然成立。因而假设检验步骤与上述大样本条件下的假设检验步骤相同，唯一的区别在于使用的标准误替换为异方差稳健标准误。

异方差稳健标准误：一个模拟

异方差的模拟

- 在white_hetero.do中，我们首先生成$x \sim \mathcal{N}(0,1)$，并生成一个$\sigma = \sqrt{x^2}$，进而生成误差项：$u|x \sim \mathcal{N}\left(0, \sigma^2\right)$，从而人为设定了一个异方差结构。
- 在以上程序中，我们首先定义了一个程序（program）：
  - 该程序会产生‘obs’个随机观测；
  - 在Stata中，reg命令后面加入选项robust就可以汇报基于异方差稳健标准误的结果，因而在该程序中加入了一个「robust」选项
    - 当使用该程序时，如果加入「robust」选项，在最小二乘估计时也会加入robust选项；如果不加该选项，则默认不适用怀特异方差稳健标准误。
- 最后我们返回了b的估计值、b的标准误以及是否拒绝了$H_0 : b = 0$的原假设。最后通过simulate前缀，重复该命令1000次，并记录下每一次b的估计值、b的标准误以及是否拒绝了原假设。
- 我们设$b = 0$运行了以上程序，经过1000次模拟

异方差稳健标准误：一个模拟

异方差的模拟

结果：

1. $\hat{b}$的均值大约为0.007，与0相差无几，说明对$\hat{b}$的估计并没有出现太大的偏差；
2. 1000次$\hat{b}$估计的标准差为0.175左右；
3. 如果不使用怀特异方差稳健的标准误，估计的s.e.$(\hat{b})$平均为$0.099 \ll 0.175$，严重低估了$\hat{b}$估计的标准误；
4. 相应的，如果不使用怀特异方差稳健的标准误，大约有25%次拒绝了原假设，远远大于5%，意味着假设检验出现了问题；
5. 如果使用异方差稳健的标准误，估计的s.e.$(\hat{b})$平均为$0.16 \approx 0.175$，与$\hat{b}$估计的标准差相差无几；
6. 相应的，使用异方差稳健的标准误的情况下，大约有6%拒绝了原假设，约等于5%，意味着假设检验没有出现严重问题。

一个实例：OHIE实验中的假设检验

OHIE实验

在美国，Medicaid是针对穷人的健康保险计划。在2008年时，俄勒冈州计划恢复Medicaid中的OHP Standard计划。由于预计申请人数非常多，因而州政府推出了一个按照抽签分配名额的方法。个人一旦被抽中，整个家庭都可以享受该计划。在个人被抽中后，州政府会联系申请人参加计划，然而由于种种原因，并非所有抽中的人最终都参加了该计划。为了检验实际数据中抽签是否仍然是随机的，Finkelstein等人(2012)使用了如下回归:

$$y_i = \beta_0 + \beta_1 \times \text{treatment}_i + u_i$$

其中treatment_i为个体i是否抽中签（=1为处理组，=0为对照组），y_i为一些人口统计指标，比如出生年份、性别、语言等等，如果$\hat{\beta}_1$不显著，那么可以认为抽中和没有抽中标签的两组人具有相似的特征。（ohie_qje_test_one.do）

一个实例：OHIE实验中的假设检验

OHIE实验

变量	(1) 出生年份	(2) 性别	(3) 英语问卷	(4) 本人签署	(5) 首日填报	(6) 有电话号码
处理组	0.253*	-0.0172***	-0.00647*	-0.0440***	0.00530	-0.00196
	(0.148)	(0.00615)	(0.00370)	(0.00438)	(0.00351)	(0.00421)
常数项	1,968***	0.556***	0.909***	0.876***	0.0900***	0.871***
	(0.106)	(0.00439)	(0.00262)	(0.00296)	(0.00247)	(0.00298)
Observations	33,823	33,822	33,823	33,823	33,823	33,823
R-squared	0.000	0.000	0.000	0.004	0.000	0.000

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

多重共线性

- 在最小二乘法的假设中，要求$\sum_{i=1}^{N} x_i x_i' = X'X$可逆，如果该假设不满足，我们称为**完美共线性（perfect collinearity）**。
  - 如果存在完美共线性，意味着变量之间可以完美的线性表示出，或者存在完美的线性相关性，那么回归方程是不可识别的。
- 然而现实的情况可能是，虽然$X'X$可逆，但是变量之间的相关性非常高，此时，留给每一个变量的**变异（variation）**可能会太小，从而导致对该系数的推断有失准确。

系数的标准误

- 考虑如下回归:

$$y_i = \gamma w_i + \tilde{x}_i' \tilde{\beta} + u_i$$

- 同方差下$\hat{\gamma}$的渐进方差的估计为:

$$\widehat{\mathbb{V}(\hat{\gamma})} = \frac{s^2}{(N-1)\, s_w^2\, (1 - R_w^2)}$$

系数的标准误和多重共线性

- 即w的系数$\hat{\gamma}$的标准误受到以下几个因素的影响：
  1. 误差项u的方差越大，则s^2越大，$\hat{\gamma}$的方差越大；
  2. 样本量N越大，则$\hat{\gamma}$的方差越小；
  3. 变量w的样本方差s_w^2越大，则$\hat{\gamma}$的方差越小；
  4. R_w^2越小，即其他变量对于x_w的预测能力越弱、相关性越小，则$\hat{\gamma}$的方差越小。
- 其中第4条意味着，即使不存在完美共线性，变量之间的高度相关也会导致估计量方差变大，我们称这种现象为**多重共线性**（multicollinearity）。
- 相比于w与其他解释变量w之间不相关的情况，$\hat{\gamma}$的方差被扩大了$\frac{1}{1-R_w^2}$倍，因而该数值也被称为**方差膨胀因子**（variance inflation factor, VIF）。

多重共线性的后果

- 多重共线性会导致一些我们不希望看到的后果，比如：
  - 由于回归系数的方差比较大，因而所估计处的系数非常不稳定，精度很差，甚至会出现本来为正，估计出来却是负的情况，或者相反；
  - 同样由于由于回归系数的方差比较大，虽然估计出来的系数比较大，但是由于标准误也非常大，假设检验的结果可能是不显著，此时更容易犯第II类错误。
    - 问题：第I类错误呢?

多重共线性的模拟

多重共线性的后果

在下面程序中，我们设定$x_1 \sim \mathcal{N}(0,1)$，并在以下两种情况：

1. $mc = 0$时，令$x_2 \sim \mathcal{N}(0,1)$
2. $mc = 1$时，令$x_2 = \frac{3}{\sqrt{10}}x_1 + \frac{1}{\sqrt{10}}e, e \sim \mathcal{N}(0,1)$

从而在第一种情况下，x_1, x_2之间不存在相关性，而在第二种情况下，x_1, x_2之间存在着极强的相关性。注意在以上两种情况下，x_2的方差都是一样的。之后使用：

$$y = b \cdot x_1 + x_2 + u, u \sim \mathcal{N}(0,1)$$

产生y，并进行回归，重复2000次（multi_colinearity.do）

我们分别选取了四种情况下，计算了在原假设$H_0 : b = 0$的条件下，拒绝原假设的比例、b估计值$\hat{b}$的平均值$\bar{\hat{b}}$、$\hat{b}$的标准差s.d. $\left(\hat{b}\right)$、$\hat{b}$的标准误的平均值s.e. $\left(\hat{b}\right)$以及估计出的系数$\hat{b}$为负数的比例neg

多重共线性的模拟

多重共线性的后果

	b	多重共线性	N	拒绝率	$\bar{\hat{b}}$	s.d. $\left(\hat{b}\right)$	$\overline{\text{s.e.}\left(\hat{b}\right)}$	neg
(1)	0	有	20	5.85%	−0.026	0.794	0.769	50.3%
(2)	1	有	20	23.6%	0.979	0.768	0.804	10.3%
(3)	1	无	20	94.65%	0.991	0.243	0.253	0.05%
(4)	1	有	200	85.65%	1.007	0.321	0.327	0.1%

多重共线性的模拟

- 在所有的4行中，所有的估计偏差都不大，且所有的标准误的估计（的均值）与$\hat{b}$的标准差都相差不大，意味着标准误的估计问题不大。
- 在(1)行中，在$b = 0$的情况下，我们可以看出犯第一类错误的概率大约为5.85%，与理论值5%非常接近，意味着假设检验的水平（size），或者犯第一类错误的概率没有出现太大问题。
- 在相同样本量下，通过比较第(2)行与第(3)行，可以看到多重共线性的存在使得s.d. $\left(\hat{b}\right)$增加了，因而也减少了检验的势（power），也就是犯第二类错误的概率提高了。
- 在相同样本量下，通过比较第(2)行与第(3)行，可以看到多重共线性的存在使得估计出的系数以更大的概率出现了符号相反的情况，意味着估计出的系数更加不稳定了。
- 比较第(2)行与第(4)行，可以发现随着样本量变大，检验的势有极大改善，同时系数符号估计反的情况也得到了极大改善。

多重共线性的后果

- 多重共线性的存在并没有像异方差问题一样，使得标准误的估计出现问题。
  - 多重共线性导致的问题是在给定样本量的条件下，估计系数的标准差变大了。
  - 由于标准误的估计没有问题，因而犯第I类错误的概率也没有被影响。
- 但是，由于标准误变大了，犯第二类错误的概率却变大了，这意味着，即使某个变量有影响，由于多重共线性的存在，却很有可能不能拒绝原假设。
- 另外，多重共线性是一个小样本问题，当N变大时，多重共线性导致的问题都可以被有效解决。
- 综合以上，多重共线性在小样本且系数不显著的情况下，可能是比较棘手的问题。

多重共线性的诊断

- 方差膨胀因子
- 另一个比较有效的方法时计算矩阵$X'X$的**条件数（condition number）**
  - 矩阵$X'X$经过标准化之后，最大特征值与最小特征值之比即条件数。
  - 条件数越大，代表变量之间的相关性越严重，共线性也越严重
  - 一般认为条件数大于20时需要关注多重共线性的问题。

多重共线性的处理

- 如果存在多重共线性，通常非常难以处理：
  - 理论上最行之有效的方法是扩大样本量，然而在观测数据中通常是不可行的。
  - 常用的其他方法，比如删除变量、主成分分析等方法，都可能带来更加严重的问题。
  - 此外，使用Lasso或者岭回归方法会使得系数不可避免地产生偏差，因而如果我们关心的是因果推断，除非在高维情况下，这些方法也应该尽量避免使用。
- 幸运的是，只有小样本且核心解释变量不显著时，才需要担心多重共线性问题。
- 在实际应用中应该尽量使用观测更多或者更加微观的数据，结论会更加稳定。

受限最小二乘

- 有时回归模型中的系数可能存在天然的约束条件，此时我们可以通过**约束最小二乘（constrained least squares）**或者**受限最小二乘（restricted least squares）**的方法得到估计量。
- 为了简单起见，我们这里讨论带有线性约束的最小二乘估计问题，其中线性约束为：

$$R\beta - q = 0$$

其中R为$r \times K$的（行满秩）矩阵，r为约束个数，q为$r \times 1$的矩阵。

线性约束举例

生产函数的约束

比如，一个可能的场景是对于生产函数：$Y = AK^{\alpha}L^{\delta}$两边取对数并加入误差项，得到：

$$\ln Y = \eta + \alpha \ln K + \delta \ln L$$

其中$\eta = \ln A$为不可观测的误差项，我们可能需要检验该生产函数是不是规模报酬不变的，这等价于检验：$\alpha + \delta = 1$。或者我们可以令：

$$\beta = \begin{pmatrix} \eta \\ \alpha \\ \delta \end{pmatrix}, R = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix}, q = 1$$

那么$\alpha + \delta = 1$可以写为：$R\beta - q = 0$

线性约束举例

多个约束

如果在回归:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i$$

如果理论上满足$\beta_1 = 0, \beta_2 + \beta_3 = 1$，那么以上两个约束可以使用:

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}, R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}, q = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

带入$R\beta + q = 0$来表示，此时$r = 2$，即有两个约束。

受限最小二乘

- 我们将$R\beta - q = 0$作为最小二乘法目标函数的约束条件，即可解得受限最小二乘估计量：

$$\min_{\beta} (Y - X\beta)' (Y - X\beta)$$
$$s.t.\ \ R\beta - q = 0$$

- 通过拉格朗日乘子法，以上最优化问题等价于最小化：

$$\mathscr{L}(\beta, \lambda) = (Y - X\beta)' (Y - X\beta) - 2\lambda' (R\beta - q)$$

其中2λ为$r \times 1$的拉格朗日乘子，因为有r个约束，相应的有r个拉格朗日乘子。

- 其一阶条件为：

$$\begin{cases} 2X'X\beta - 2X'Y - 2R'\lambda = 0 \\ 2(R\beta - q) = 0 \end{cases}$$

受限最小二乘

- 将第一个式子带入到第二个式子中，得到：

$$R\left[\left(X'X\right)^{-1}\left(X'Y+R'\lambda\right)\right]-q=0$$

即：

$$\lambda=\left[R\left(X'X\right)^{-1}R'\right]^{-1}\left(q-R\hat{\beta}\right)$$

其中$\hat{\beta}$为不受限的普通最小二乘估计量。

- 将上式带回到第一个式子中，就可以解得：

$$\hat{\beta}_c=\hat{\beta}-\left(X'X\right)^{-1}R'\left[R\left(X'X\right)^{-1}R'\right]^{-1}\left(R\hat{\beta}-q\right)$$

以上便是在约束$R\beta-q=0$的条件下得到的最小二乘估计量。

联合检验

- 以上我们介绍了单个系数的假设检验问题，然而在许多应用中，我们需要对多个系数进行联合检验
- 对回归方程：

$$y_i = x_i'\beta + u_i$$

我们希望检验在真实的数据生成过程中:$R\beta - q = 0$是否成立，即检验原假设：

$$H_0 : R\beta - q = 0$$

以及备择假设$R\beta - q \neq 0$。

联合检验的构造

- 为了构造检验统计量，我们可以分别讨论原假设成立以及备择假设成立条件下不受限的最小二乘估计量$\hat{\beta}$以及首先最小二乘估计量$\hat{\beta}_c$之间的差异。
- 如果原假设成立，即$R\beta - q = 0$成立，如果记$\hat{u}$为无约束下最小二乘估计量$\hat{\beta}$计算的残差，$\hat{u}_c$为受限最小二乘估计量计算得到的残差，那么如果原假设成立，应当有$\text{RSS}_c - \text{RSS} \approx 0$，其中

$$\text{RSS}_c = \sum_{i=1}^{N} \hat{u}_{c,i}^2,\ \text{RSS} = \sum_{i=1}^{N} \hat{u}_i^2$$

- 然而当原假设不成立时，$\hat{\beta}_c \neq \hat{\beta}$，此时两个残差平方和应该有非常大的差异，或者$\text{RSS}_c \gg \text{RSS}$。

F统计量

- 为了进行假设检验，我们必须推导出在原假设成立的条件下，$\mathrm{RSS}_c - \mathrm{RSS}$的分布情况。
- 注意到以上的残差之差一定依赖于误差项的方差，所以为了构造一个基准（pivotal）统计量，我们还可以继续将以上的残差平方和除以误差项方差的估计，即$s^2 = \mathrm{RSS}/(N-K)$。
- 为此，如果假设$u|x \sim \mathcal{N}\left(0, \sigma^2 I\right)$，假设原假设$R\beta - q = 0$成立，我们有如下结论：

$$\frac{(\mathrm{RSS}_c - \mathrm{RSS})/r}{\mathrm{RSS}/(N-K)} \sim \mathrm{F}\left(r, N-K\right)$$

（证明见讲义）

F检验

- 而如果备择假设成立，显然$\text{RSS}_c - \text{RSS}$应该显著大于0。
- 为此，我们可以使用以上定理结论，通过计算：$F = \frac{(\text{RSS}_c - \text{RSS})/r}{\text{RSS}/(N-K)}$使用右侧检验：
  - 给定显著性水平α，当$F > \text{F}_\alpha^{(r,N-K)}$时，即拒绝原假设，
    - 其中$\text{F}_\alpha^{(r,N-K)}$为$\text{F}(r, N-K)$分布的$(1-\alpha)$分位数，即右侧$\alpha$的临界值。

Stata中的F检验

- F检验在Stata中可以在回归结束后直接使用“test”命令完成。
- 比如在OHIE的例子中，每次回归阶数后我们都用“test treatment”进行了对单个参数的F检验，其原假设为：$H_0 : \beta_{\text{treatment}} = 0$。
- 实际上，使用以上的F检验和单个系数的t检验是完全等价的（练习题）。

F检验：Stata实例

NTV与俄罗斯选举

Enikilopov、Petrova和Zhuravskaya（2011）使用俄罗斯大选的数据检验了媒体对于选举的影响，具体的，他们讨论了独立国家电台NTV的接收情况对于选举的影响，不过在此之前，需要讨论NTV接收情况是否受到不同地区政治偏好的影响，为此，他们做了如下回归：

$$\text{NTV}_{s,1999} = \alpha + \sum_{j=1}^{J} \delta_j \text{Vote}_{sj,1995} + x_s'\beta + u$$

其中$\text{NTV}_{s,1999}$为选区s在1999年时NTV的接收情况，$\text{Vote}_{sj,1995}$为选区s在1995年时对于各个政党的支持率以及总的投票率（turnout），x_s为其他的社会经济变量、地区虚拟变量（每个地区有多个选区）等。为了说明政治偏好对于NTV的设立没有影响，需要联合检验：

$$H_0 : \delta_j = 0, j = 1, 2, ..., J$$

（test_jointly_ntv.do）

拟合优度检验

- 在线性回归的预测和拟合中，可以使用可决系数R^2度量拟合优度
- 更进一步的，我们有时还希望对模型的拟合优度进行检验。
- 最常见的检验是，当使用x对y进行预测时，其预测效果是否比仅仅使用平均值$\bar{y}$对y进行预测究竟更好，或者x是否为对y的预测带来了更多的信息。
- 在线性回归的情景下，该检验相当于检验是否所有的β_j（除常数项）都同时等于0。如果设截距项为β_1，那么我们需要检验的原假设为：

$$H_0 : \beta_2 = \beta_3 = \cdots = \beta_K = 0$$

拟合优度检验

- 原假设为：$H_0 : \beta_2 = \beta_3 = \cdots = \beta_K = 0$
- 对于以上假设检验的问题，我们可以使用上一节所提到的F检验：
  - 在这里，原假设包含了$r = K - 1$个假设。
  - 如果H_0成立，意味着我们只是使用了常数项对y进行拟合，因而在受限最小二乘估计中，常数项的估计值：$\hat{\beta}_{1c} = \bar{y}$，而其他系数的估计值为$\hat{\beta}_{kc} = 0, k = 2, ..., K$，因而受限最小二乘的残差平方和为：

$$\mathrm{RSS}_c = \sum_{i=1}^{N} (y_i - \bar{y})^2 = Y' M_0 Y = \mathrm{TSS}$$

  - 而非受限最小二乘的残差平方和为：$\hat{u}'\hat{u} = Y'MY = RSS$
  - 因而检验统计量：

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/r}{\mathrm{RSS}/(N-K)} = \frac{\mathrm{ESS}/(K-1)}{\mathrm{RSS}/(N-K)} = \frac{R^2/(K-1)}{(1-R^2)/(N-K)}$$

  如果H_0成立，则$F \sim \mathrm{F}(K-1, N-K)$。

方差分析

- **方差分析**（Analysis of Variance, ANOVA）是在实验数据中常用的统计方法之一。
- 在实验中，通常将实验个体分为很多个组进行实验，并观察每个组的结果变量y是否有显著的差异。
- 在该方法中，我们将实验中的分类变量称之为**因子**（factor），而分类变量的取值称之为**水平**（level）
  - 比如，在粮食优惠券的实验（Jensen和Miller，2011）中，作者将低收入家庭分为了四组：不给优惠券；给予0.1元、0.2元、0.3元每斤的优惠券
  - 在这里，因子即家庭持有何种优惠券，而水平分为0元、0.1元、0.2元、0.3元每斤四种
  - 方差分析的目的即研究这四组的结果变量（比如营养摄入）是否有显著的不同。

方差分析与线性回归

- 方差分析可以看作是回归分析的一种特例。
- 比如，在以上例子中，由于我们现在把优惠券状态分为4组，将其看为分组变量而非数值变量，因而我们可以使用虚拟变量回归：

$$y = \beta_1 + \beta_2 d_2 + \beta_3 d_3 + \beta_4 d_4 + u$$

其中$d_k, k = 1, ..., 4$为家庭组别的虚拟变量。

- 根据之前的结论，有：

$$\begin{cases} \beta_1 = \mathbb{E}(y|d = 1) \\ \beta_k = \mathbb{E}(y|d = k) - \mathbb{E}(y|d = 1) \quad k = 2, 3, 4 \end{cases}$$

因而检验每个分组的均值都相等，等价于检验：$H_0 : \beta_2 = \beta_3 = \beta_4 = 0$。

方差分析与线性回归

- 一般地，对于一个分组变量$G_i \in \{1, 2, ..., K\}$，即K个分组，令$d_{ik} = \mathbb{1}\{G_i = k\}$为虚拟变量，可以使用回归：

$$y_{ik} = \beta_1 + \sum_{k=2}^{K} \beta_k d_{ik} + u_{ik}$$

对β_k进行估计，并使用F检验对原假设$H_0 : \beta_k = 0, k = 2, ..., K$，即每个组别的均值都相等，进行检验。

- 根据上一节的结论，以上检验的检验统计量可以写为：

$$F = \frac{\mathrm{ESS}/(K-1)}{\mathrm{RSS}/(N-K)}$$

在原假设的条件下$F \sim \mathrm{F}(K-1, N-K)$。

方差分析

- 注意到，在上述最小二乘回归中，由于只有虚拟变量作为解释变量，因而对于y_{ik}的最优预测即其所在组别的均值：如果$d_{ik}=1$，那么$\hat{y}_{ik}=\bar{y}_k$，因而：

$$\begin{cases}\text{TSS}=\sum_{i=1}^{N}\left(y_{ik}-\bar{y}\right)^2=Y'M_0Y\\ \text{RSS}=\sum_{i=1}^{N}\sum_{k=1}^{K}\left[d_{ik}\left(y_{ik}-\bar{y}_k\right)^2\right]\\ \text{ESS}=\sum_{i=1}^{N}\sum_{k=1}^{K}\left[d_{ik}\left(\bar{y}_k-\bar{y}\right)^2\right]=\sum_{i=1}^{N}\left[N_k\left(\bar{y}_k-\bar{y}\right)^2\right]\end{cases}$$

- 由于RSS为y_{ik}减去相应组别均值$\bar{y}_k$的平方和，因而通常也被称为**组内平方和（sum of squares within groups）**，而ESS度量了不同组别之间均值之间的差异，因而也被称为**组间平方和（sum of squares between groups）**。

方差分析举例

不同教育水平的收入

- 作为示例，我们使用CFPS数据检验不同教育水平的对数收入均值是否相等（oneway_anova.do）
- 可以看到te4变量共有7个分组，$K = 7$，样本量为$N = 3226$。使用最小二乘法的F检验中，$F = 4.42$，残差平方和为RSS $= 72794.67$；使用方差分析得到的结果中，$F = 4.42$, RSS $= 72764.67$，结果完全相同。

方差分析：更多比较

当然，我们也可以使用F检验单独比较某两个、三个组之间的均值是否相等，或者其他任意线性原假设。比如：

不同教育水平的收入

- 使用单个参数的t检验，比如使用文盲/半文盲（te4=1）作为基准组，并检验小学学历（te4=2）前面的系数是否为0：

```
reg log_income ib1.te4
```

- 使用F检验直接检验小学学历（te4=2）前面的系数等于0：

```
reg log_income ib1.te4
test 2.te4=0
```

- 使用所有的虚拟变量，并删掉常数项的回归，并直接检验$H_0 : \beta_1 = \beta_2$：

```
reg log_income ibn.te4, noconstant
test 1.te4=2.te4
```

结构变化检验

- 有时我们会关心两个或者多个组别的分组回归系数是否相等。
  - 比如，我们可能关心某个时间节点前后，同一个线性回归的回归系数是否完全相等
  - 或者我们会关心不同性别的回归方程是否完全相等。
- 不失一般性，我们令$d_i = 0/1$代表一个分类变量，我们可以根据d_i进行分组，并分别在组内进行回归：

$$y_i = x_i'\beta_0 + u | d_i = 0$$
$$y_i = x_i'\beta_1 + u | d_i = 1$$

  其中β_0, β_1, x_i为K维向量，我们希望检验$H_0 : \beta_0 = \beta_1$。
- 然而，以上两个回归系数分别在两个回归方程中，所以无法直接进行比较检验。

结构变化检验

- 解决以上问题的办法是想办法将其转化为一个回归。实际上，以上回归等价于：

$$y_i = (d_i \cdot x_i)' \beta_1 + [(1 - d_i) \cdot x_i]' \beta_0 + u_i$$

即将x的每个分量（包括常数项）都乘以d_i和$1 - d_i$进行回归。

- 因而，为了检验原假设：$H_0 : \beta_0 = \beta_1$，可以直接估计上式，并使用F检验即可。
- 注意由于β_0, β_1均为K维向量，因而在该原假设中，有$2K$个参数以及K个约束，因而F分布应为$\mathrm{F}(K, N - 2K)$。以上检验由邹至庄于1960年提出（Chow, 1960），因而也被称为**邹氏检验（Chow test）**。

邹氏检验实例

比较不同城市回归

我们使用如下回归方程比较一二线城市（北上广深+省会城市）与其他城市的人口增长模式是否有不同:

$$y_i = \beta_{01}(1-d_i) + \beta_{11}d + \beta_{02}(1-d_i)\cdot x_{i2} + \beta_{12}d\cdot x_{i2} + \beta_{03}(1-d_i)\cdot x_{i3} + \beta_{13}d_i\cdot x_{i3} + u_i$$

其中y为城市的人口自然增长率，x_2为人口对数，x_3为人口密度对数，该比较的原假设为:

$$H_0: \begin{cases} \beta_{01} = \beta_{11} \\ \beta_{02} = \beta_{12} \\ \beta_{03} = \beta_{13} \end{cases}$$

chow_test.do

邹氏检验

- 或者，以上检验式也可以写为：

$$y_i = x_i'\beta_0 + d_i \cdot x_i'(\beta_1 - \beta_0) + u \triangleq x_i'\beta_0 + d_i \cdot x_i'\delta + u_i$$

其中$\delta = \beta_1 - \beta_0$

- 从而对原假设：$H_0 : \delta = 0$进行检验也可以达到同样的目的
- 比如，使用该思路，上例的检验可以通过如下回归方程完成：

$$y_i = \gamma_1 + \gamma_2 x_{i2} + \gamma_3 x_{i3} + \delta_1 d_i + \delta_2 d_i \cdot x_{i2} + \delta_3 d_i \cdot x_{i3} + u_i$$

我们需要检验原假设：$H_0 : \delta_1 = \delta_2 = \delta_3 = 0$：

```
reg pop_growth log_pop log_pop_dens d d_log_pop d_log_pop_dens
test d d_log_pop d_log_pop_dens
```

分组系数检验

- 有时，我们可能仅仅对某一个变量在两个组别之间的差异，也可以使用如上思路
  - 比如，如果仅仅对两种不同城市的x_2对人口自然增长率的影响，只需要在以上回归中检查d_log_pop的系数是否显著即可。
- 或者，有时也会假设其他变量对y的影响在两个组别没有区别，从而只需要加入虚拟变量d、d和关心的解释变量的交乘项即可。
  - 比如如果我们关心的是x_2的系数在两种不同城市之间的系数是否不同，也可以使用回归:

$$y_i = \gamma_1 + \gamma_2 x_{i2} + \gamma_3 x_{i3} + \delta d + \delta_2 d \cdot x_{i2} + u$$

  然后检验$H_0 : \delta_2 = 0$即可。
  - 注意在以上回归方程中没有包含$d \cdot x_3$，即假设了x_3的系数在两个组别是相同的。
  - 此外因为加入了$d \cdot x_2$，所以一般虚拟变量d也需要单独加入在回归方程中。