第三章、线性回归模型的应用

- 3．1．对单个参数的假设检验 t-检验
- 3.3.1 t-检验
- 对多元回归模型中的任一参数进行假设检验
- 定理3．1（OLS估计的t分布）在CLM假设1~6之下，$(\hat{\beta}_j - \beta_j)/se(\hat{\beta}_j) \sim t_{n-k-1}$ 。其中k为模型中未知斜率参数个数或自变量个数
- 大部分情况下，我们想要检验的零假设都是（null hypothesis） $H_0 : \beta_j = 0$
- 零假设就表示考虑了 $x_1, \cdots, x_{j-1}, x_{j+1}, \cdots, x_k$ 之后，x_j 的取值对y的期望值没有什么影响

1

- 回归方程

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 \exp er + \beta_3 tenure + \varepsilon$$

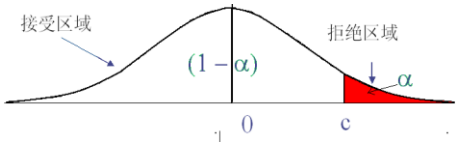
- 零假设 β_3 =0就表示考虑教育程度和工龄之后，tenure 对一个人的工资水平没有影响
- 对上面的零假设进行检验的统计量通常是t统计量定义为 $t_{\hat{\beta}_j} = \hat{\beta}_j / se(\hat{\beta}_j)$
- $\hat{\beta}_j$ 离开零点越远就越能给出反对 H_0 的证据
- $\hat{\beta}_j$ 必须要用样本的误差来进行加权调整
- t统计量反映了 $\hat{\beta}_j$ 离开零点多少个标准差

2

- 严格的拒绝规则依赖于备择假设
- 所选取的显著水平(significance level)
- 显著水平就是，当 H_0 成立时而拒绝了它的概率
- 零假设成立时 $t_{\hat{\beta}_j}$ 的样本分布正好是
- 自由度为n-k-1的t分布
- 3．1．2单边备择假设的检验 $H_1 : \beta_j > 0$
- 由于某种经济理论或其它原因使我们确定不会比零小
- 可以认为其它零假设就是 $H_0 : \beta_j \le 0$

3

- 希望找到足够大的、为正的 $t_{\hat{\beta}_j}$ 的取值来拒绝 H_0 而接受 H_1
- 5%显著水平下足够大的，就是超过自由度为n-k-1的t分布的95%分位点，记为c
- 决定是否拒绝的数值c称为临界值



4

- 在5%的显著水平下，自由度为n-k-1=30的临界值为c=1.697
- 10%显著水平下，自由度为40的临界值c=1.303
- 5%的显著水平下，自由度为120的t分布临界值为1.658，正态分布为1.645
- 例3.1 工资与教育程度的估计模型为

$$\log(\hat{wage}) = 0.284 + 0.092 educ + 0.0041 \exp er + 0.022 tenure$$

（0.104） (0.007) (0.0017) (0.03)

零假设 $H_0 : \beta_{exper} = 0$ 备择假设为 $H_1 : \beta_{exper} > 0$

5

自由度为526-3-1=522，在1%显著水平下的临界值为2.326，t统计量为 $t_{\hat{\beta}_{exper}} = 0.0041/0.0017 \approx 2.41$

拒绝工龄对工资没有影响的零假设

备择假设为小于0的单边假设拒绝规则正好是前一情形的镜像 ,拒绝规则为 $t_{\hat{\beta}_j} < -c$

3．1．3 双边备择假设 $H_1 : \beta_j \neq 0$

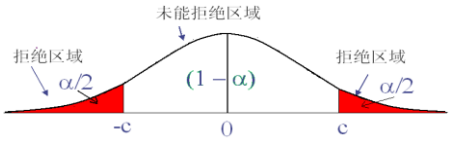
希望在其它条件相同时 x_j 对y有偏效应，而不论是正的还是负的效应。

没有很好的相关理论来确保参数的符号

备择假设是双边时，可用绝对值来表示t统计量

6

- 拒绝规则为 $|t_{\hat{\beta}_j}|>c$
- c为选取的临界值
- 5%显著水平，对双边检验，所选择的c是要使t分布在每一边各自有2.5%的概率
- 要取自由度为n-k-1的t分布的97.5%分位点



7

- 当n-k-1=30时，5%显著水平双边检验的临界值为c=2.042
- 如果 H_0 在 $\alpha\%$ 的显著水平下被拒绝，通常说：H_0 是在 $\alpha\%$ 显著水平下统计显著的，或统计意义下异于零。若没有拒绝 H_0，就说：是在 $\alpha\%$ 显著水平下统计不显著

- 3．1．4　其他备择假设
- 有时我们也会遇到要检验是否等于其它给定常数的假设
- 把零假设表示为 $H_0:\beta_j=a_j$

8

- 检验该零假设的t统计量为 $t=(\hat{\beta}_j-a_j)/se(\hat{\beta}_j)$
- 其他方面与前面的假设检验一样
- 例零假设和备择假设分别为

$$H_0:\beta_j=1 \qquad H_1:\beta_j>1$$

- 一样的临界值，差别只在于我们计算的t统计量不一样

9

3．1．5计算t检验的p-值（p-value）

- 经典假设检验有活动的变数
- 事先选定一个显著水平
- 没有一个公认为合适的显著水平
- 可能会隐藏一部分假设检验的信息 。
- 例双边备择假设，自由度为40时，得到t统计量的计算值为1.85，5%的水平下不能拒绝零假设，在10%水平拒绝零假设的结论
- 给定的t统计量计算值零假设被拒绝的最小的显著水平 ？ 这一水平就是这一检验的p-值

10

- 例子中的p值为0.0718
- p-值是一个概率，所以它总在0和1之间
- 需要有比较详细的t分布表或采用程序来计算
- 在 $H_0:\beta_j=0$　对双边备择假设时，p-值就是概率 $P(|T|>|t|$
- p-值是在零假设为真时，我们所观测到的t统计量值所能达到的最大概率
- p-值越小，就越有证据说明零假设不成立
- 例如果零假设成立，我们有7.2%的概率会观测到t统计量的绝对值大于1.85

11

- 3．1.6 经济或实际与统计显著性
- 自变量的统计显著性主要取决于 $t_{\hat{\beta}_j}=\hat{\beta}_j/se(\hat{\beta}_j)$
- 经济意义的显著性是看 β_j 的大小
- 统计显著时，可能是由于 β_j 足够大，也可能是由于 $se(\hat{\beta}_j)$ 足够小
- 正确区分t统计量显著的原因是非常重要的
- 例3.2 。考虑公司的销售收入增长与年薪最高的前三位CEO的年薪与公司特征之间关系。lnthreeceo为年薪最高三位CEO的年薪之和取对数，lnasset为公司总资产取对数，lnincome为公司销售收入取对数，eachearn为每股收益，herfindahl为反映公司前10大股东分散程度的Herfindahl指数

12

$$\widehat{\ln threeceo} = 8.394 + 0.181 \cdot \ln asset + 0.053 \cdot \ln income - 0.625 herfindhal$$

（0.324） （0.018） （0.011） （0.205）

lnincome的t统计量为t=-0.0053/0.011=-4.67

结果企业的销售收入增加1%人，最高三位CEO的年薪总和只会增加0.05个百分点。

根据样本量来选择合适的显著水

样本为数百水平时，选择5%的水平

当有数千样本时，用1%的水平

样本量已经不小了，某些模型参数的估计方差比较大可能也是因为有多重共线性的影响

13

- 经济显著性和统计显著性可以归结为三点
  - 首先检查统计显著性，是统计显著的，再考虑该变量在经济意义下的重要程度，需要认真考虑因变量和自变量的关系。
  - 统计上不显著，需要考虑该变量对y是否有预计的影响，影响在实际中足够重要；给出p-值。对比较小的样本，有时给出的p-值可以达到0.2；但这样的结果是比较弱的，处理时要多加小心。
  - 碰到变量虽然有比较小的t统计量值，但却给出了相反的符号。应该显著的变量出现了非预期的符号和比实际有更强的影响是非常麻烦的问题

14

- 3．1.7 置信区间
- 置信区间（confidence interval，CI）也称为区间估计，提供了真实参数可能取值的一个范围的估计
- 未知参数 β_j 的95%置信区间为 $\hat{\beta}_j \pm c \cdot se(\hat{\beta}_j)$
- 置信区间的意思是：当我们用随机样本反复对模型的系数进行估计时，我们会有95%的样本给出模型的真实参数会在上、下限之间
- 例一个模型，其自由度为n-k-1=25，对任意参数一个95%的置信区间为

$$[\hat{\beta}_j - 2.06 \cdot se(\hat{\beta}_j), \hat{\beta}_j + 2.06 \cdot se(\hat{\beta}_j)]$$

15

- 3．1.8 对参数线性组合的检验
- 对多个参数的一个线性组合进行检验
- 考虑大学教育和专科教育，记jc为接受两年的专科教育，univ为接受4年的本科教育，所选的样本都具有高中毕业学历。模型为

$$\log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 \exp er + \varepsilon$$

- 接受1年的专科教育是否与接受1年的本科教育相同，即零假设为 $H_0 : \beta_1 = \beta_2$
- 备择假设为单边的，即1年的专科教育不如1年的本科教育 $H_1 : \beta_1 < \beta_2$
- 改写假设为 $H_0 : \beta_1 - \beta_2 = 0 \quad H_1 : \beta_1 - \beta_2 < 0$

16

- 检验的t统计量 $t = (\hat{\beta}_1 - \hat{\beta}_2) / se(\hat{\beta}_1 - \hat{\beta}_2)$
- 可以得到相应t分布的自由度和选定的置信水平来给出检验，给出临界值、p-值等
- 从模型估计得到

$$\widehat{\log(wage)} = 1.472 + 0.0667 jc + 0.0769 univ + 0.049 \exp er$$

（0.021） （0.0068） （0.0023） （0.0002）

- 两个估计的差为 $\hat{\beta}_1 - \hat{\beta}_2 = 0.0102$
- 但没有给出信息来直接得到 $\hat{\beta}_1 - \hat{\beta}_2$ 的标准差
- 首先要得到它们的方差

$$Var(\hat{\beta}_1 - \hat{\beta}_2) = Var(\hat{\beta}_1) + Var(\hat{\beta}_2) - 2Cov(\hat{\beta}_1, \hat{\beta}_2)$$

17

$$se(\hat{\beta}_1 - \hat{\beta}_2) = \{[se(\hat{\beta}_1)]^2 + [se(\hat{\beta}_2)]^2 - 2s_{12}\}^{1/2}$$

- 另一方法通过调整参数而从另一模型里直接得到
- 定义这两个参数之差为 $\theta = \beta_1 - \beta_2$
- 代入原来的模型就得到

$$\log(wage) = \beta_0 + (\theta + \beta_2) jc + \beta_2 univ + \beta_3 \exp er + \varepsilon$$

$$= \beta_0 + \theta \cdot jc + \beta_2 (jc + univ) + \beta_3 \exp er + \varepsilon$$

例中合成变量正好表示接受高等教育的年数totcoll

$$\widehat{\log(wage)} = 1.472 - 0.0102 jc + 0.0769 totcoll + 0.0049 \exp er$$

（0.021） （0.0069） （0.0023） （0.0002）

计算出t统计量为-0.0102/0.0069=-1.48，p值大约为0.07

18

3．2 检验多个线性约束的F检验

- 3．2．1．多个线性约束的F检验
- 如何同时对多个参数的约束进行假设检验
- 检验一组自变量对y的偏效应都为零的情形
- 假定一个具有k个自变量的模型

$$y=\beta_0+\beta_1x_1+\cdots+\beta_kx_k+\varepsilon$$

- 假定其中的q个自变量在控制了其它变量后对y没有偏效应，零假设可以表示为

$$H_0:\beta_{k-q+1}=0,\cdots,\beta_k=0$$

- 对每个自变量作t检验会得到q个不同的拒绝规则；可能其中一些被拒绝，而另一些不能被拒绝

- 每个变量都不能被拒绝，综合起来的作用可能是不能被忽略的
- 需要作为一个整体看待，而进行联合假设检验
- 备择假设就是至少有其中之一是显著异于0的
- 回归模型的残差平方和可能是一个比较合适的综合的整体度量指标
- 分别计算完全模型和限制模型，看限制模型的残差平方和增加了多少
- 通过比较这两个模型的残差平方和来给出我们的假设检验

- 平方和之比的统计分布为F分布，称残差平方和之比所定义的统计量为F统计量

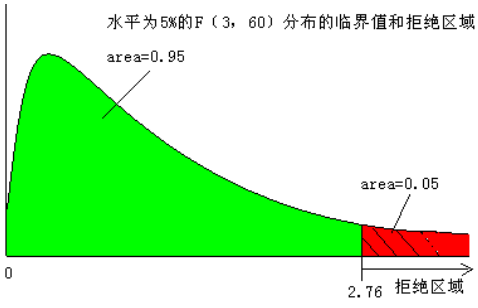
$$F=\frac{(SSR_u-SSR_\varepsilon)/q}{SSR_\varepsilon/(n-k-1)}$$

- 其中 SSR_u 为限制模型给出的残差平方和，SSR_ε 为完全模型给出的残差平方和
- F统计量的取值为非负的
- 限制模型中加入了q个约束
- 完全模型的自由度n-k-1
- 分母正好是完全模型干扰项方差的无偏估计

- 在零假设和CLM假设成立的条件下 F统计量是一个自由度为（q, n-k-1）的中心F分布，记为

$$F\sim F_{q,n-k-1}$$

- F统计量是两个 χ^2 随机变量分别除以它们的自由度之后的比值
- 当F统计量足够大时，拒绝零假设而接受备择假设
- 选取5%的水平，假定c是 $F_{q,n-k-1}$ 分布的95%分位点，这个临界值c依赖分子的自由度q和分母的自由度
- 例在5%的置信水平下，q=3和n-k-1=60的临界值c=2.76。同样自由度在1%置信水平的临界值为4.13



- 分子上的自由度通常比分母上的小
- 分母自由度超过120后F分布的变化很微小
- 当 H_0 被拒绝的时候称为联合统计显著
- 不能认为具体的某个自变量对y是否有偏效应
- 可能都有或者只有其中之一对y有影响
- 例3.3 CEO年薪例子中，只考虑公司的盈利和股东持股情况回归模型为

$$\log(threeceo)=13.09+0.0590netasset+0.0119opincom+0.3830eachearnop$$
$$(0.1445)\quad(0.0132)\quad(0.0032)\quad(0.0486)$$
$$-0.7727firsthold+0.4521hold2\_5+0.7591herfindahl5$$
$$(0.6909)\quad(0.2960)\quad(0.8544)$$
$$SSR_\varepsilon=629.08,R^2=0.1837$$

- 零假设为 $H_0:\beta_4=0,\beta_5=0,\beta_6=0$

- 在5%的置信水平下，β_3,β_4,β_5都不显著异于零，采用同样的数据对限制模型进行估计得

$$\log(threeeco) = 13.02 + 0.0534 netasset + 0.0115 opincom + 0.3894 eachearnop$$

（0.1445） (0.0132） （0.0032） (0.0481)

$$SSR_u = 635.90, R^2 = 0.1788$$

分子自由度为3，分母自由度为n-k-1=1055-5-1=1049，$F_{3,1049}$在2.5%置信水平的临界值为3.12，在1%置信水平的临界值为3.78

计算出F统计量

$$F = [(635.898 - 629.08)/3]/(629.08/1049) \approx 3.79$$

有充分的证据拒绝这3个变量没有影响的零假设

25

3．2．2．F统计量和t统计量之间的关系

结果明显与前面分别使用t检验得出的结论有冲突

自变量之间有比较高的相关性，由于多重共线性的影响，使得很难给出每个变量各自的偏效应

多重共线性对F统计量假设检验的影响不大

应用中常见变量之间很可能存在比较高的相关性

取q=1，F统计量也能对单个自变量进行显著性检验

它们之间存在关系t^2_{n-k-1}和$F_{1,n-k-1}$具有相同的分布

对一个自变量的显著性进行检验时，t统计量更加灵活，能对单边备择假设给出结论

26

- 可能一个变量在t检验下具有显著性，但一组变量的联合假设检验有没有显著性

F统计量不是回答具体某个统计量为零的最好检验

- 回归结果的达不严谨可能会混淆或掩藏了某个统计显著的变量于一组不显著的变量之中
- 例，研究一个城市的贷款批准率
- 当加入种族和年龄变量后，在5%的水平下它们联合不显著
- 一般情况下，当一个变量是非常显著时，它与其它变量的联合假设通常也会是联合显著

27

- 3．2．3． F统计量可用限制模型和完全模型的R方来计算 $SSR_u = SST(1-R_u^2)$

$$SSR_\varepsilon = SST(1-R_\varepsilon^2)$$

- F统计量可以表示为 $F = \dfrac{(R_\varepsilon^2 - R_u^2)/q}{(1-R_\varepsilon^2)/(n-k-1)}$
- 这一形式很容易给出剔除变量的F检验
- F检验的p值，与t检验类似
- F检验的p值定义为 p=P(ƒ>F)
- ƒ表示服从自由度为（q,n-k-1）的F分布随机变量
- 小的p值是反对零假设的有力证据

28

- 3．2．4 F 统计量可用来刻划一个模型的整体显著性，零假设表示为 $H_0: x_1,\cdots,x_k$ 对y没有任何解释能力，或者 $H_0: \beta_1 = \cdots = \beta_k = 0$
- 这一零假设作为限制给出的模型就不用回归，限制模型也可以表示为 $y = \beta_0 + u$
- 限制模型的R方是0，所以检验的F统计量为

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

- 这个F只能对检验所有变量是不是要整体剔除
- 例有效市场的研究中，有效市场假设就是认为过去的信息对未来市场的价格没有预测性，寻找一些变量能给出F有显著性的结果

29

- 3．2．5 用F统计量检验多个线性约束
- 例，据有关理论有一个模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

- y为资产的收益，自变量分别为市场指数收益，行业指数，地区指数，债券指数的多因素模型
- 简单市场模型认为（零假设为 ）

$$H_0: \beta_1 = 1, \beta_3 = 0, \beta_4 = 0$$

- 相应的限制模型为 $y = \beta_0 + x_1 + u + \beta_2 x_2$

$$y - x = \beta_0 + \beta_2 x_2 + u$$

- 因变量表示的是简单市场调整后的收益

30

- 分别对两个模型进行估计，而分别得到它们残差平方和为 SSR_r SSR_u
- 此时不能再使用R方来计算F统计量，因为两个方程的因变量不相同
- F统计量为 $[(SSR_u - SSR_r)/SSR_r]\cdot[(n-5)/3]$
- **3．2．6．如何报告回归的结果**
- OLS估计的系数是必不可少的，对主要的分析变量要给出单位，还需要对它进行讨论
- 估计系数的标准差或t值
- 零假设不一定就是模型的真实参数是否为0

31

- 回归的R方，模型形式分别给出来
- 同一因变量的几种自变量组合或同一模型对几组样本进行估计，最好是用一张表格进行展示
- 因变量要有明确的解释，放在表格的第一行；自变量放在第一列，标准差或t统计量用括号
- 小数后通常为三位；t值使用两为小数即可，有的结果需给出p值，通保留小数点后三位
- 注意系数的有效位数
- 显著的结果通常需要标注
- 重要的变量尽量靠前列示

32

回归自变量	事件超额收益		
第一大股东变更（dummy）	4.38 (1.95)***		
CEO变更（dummy）	6.03 (2.23)**	5.11 (1.95)***	
股权性质（dummy）	1.41 (0.49)	0.39 （0.14）	
转让比例（%）		61.49 (3.19)*	73.57 (3.77)*
转让比例平方项		-73.51 (-2.60)*	-74.36 (-2.66)*
事件窗前收益率	-10.71 (-3.46)*	-10.86 (-3.54)*	
转让价格（元）	1.20*** (1.78)	1.27*** (1.87)	
关联交易 (dummy)	2.11 (0.69)		

33

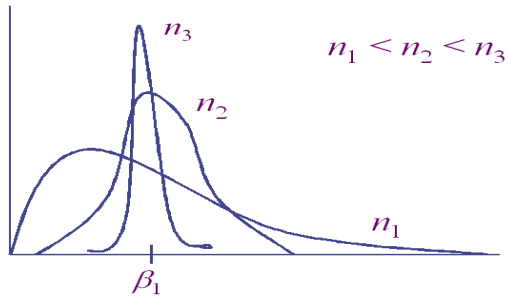
转让比例*CEO变更			15.53 (2.01)**
转让比例*股权性质			-9.60 (-1.08)
转让比例*关联交易			0.03.65 (0.43)
总资产(10亿)	-1.71 (-2.87)*	-1.52 (-2.59)*	-1.25 (-2.03)**
资产收益率	-20.08 (-1.55)	-16.13 (-1.25)	-15.10 (-1.19)
负债率	-2.63 (-0.41)	-2.73 (-0.43)	-3.29 (-0.51)
截距	-2.05 (-0.20)	-6.30 (-0.60)	-5.95 (-0.57)
R^2（%）	12.93	14.11	15.06
样本量	587	587	587

34

3．3． OLS估计的渐进性

- 假设6中干扰项与自变量独立且服从正态分布是关键
- 误差项不是正态则检验统计量也不再服从t或F分布
- 想知道样本量充分大时估计量和检验统计量的特性
- 计量学家广泛认可的标准（或最低要求）是一致性
- 如果 $\hat{\beta}_j$ 是一致估计，则随着样本量n的增加，其分布密度会越来越集中在 β_j 周围
- 样本数量的增加就能逐渐接近所关注的目标
- **3．3．1．OLS估计的一致性（consistency）**

35



36

- 概率极限的定义
- 假定$\hat{\theta}_n$是使用样本量为n的一组样本得到的关于参数的一个估计，$\hat{\theta}_n$称为是参数θ的一致估计(consistent estimetor)
- 如果对任意给定的 $\varepsilon > 0$ ，都有

$$P(|\hat{\theta}_n - \theta| > \varepsilon) \xrightarrow{n\to\infty} 0$$

- 也称 θ 是序列{ $\hat{\theta}_n$ }的概率极限，记为

$$P\lim(\hat{\theta}_n) = \theta, (n \to \infty)$$

37

- 定理3．2 （OSL估计的一致性）在假设1~4之下，OLS估计 $\hat{\beta}_j$ 是一致估计，j=0, 1, …，k

$$\hat{\beta}_1 = [\sum_{i=1}^n (x_{i1} - \bar{x}_i) y_i] / [\sum_{i=1}^n (x_{i1} - \bar{x}_i)^2]$$

$$= \beta_1 + [n^{-1}\sum_{i=1}^n (x_{i1} - \bar{x}_i)\varepsilon_i] / [n^{-1}\sum_{i=1}^n (x_{i1} - \bar{x}_i)^2]$$

- 分子分母分别取极限得到 $Cov(x_1, \varepsilon)$ 和$Var(x_1)$
- 假设4要求 $Var(x_1) \neq 0$ ，假设3有 $Cov(x_1, \varepsilon)$
- 概率极限的性质，对任意的连续函数有

$$P\lim h(\hat{\theta}_n) = h(p\lim \hat{\theta}_n), (n \to \infty)$$

$$p\lim \hat{\beta}_1 = \beta_1 + Cov(x_1, \varepsilon)/Var(x_1) = \beta_1$$

- OLS满足一致性只需要自变量与干扰项不相关

38

- 假设3’（0均值和0协方差）$E(\varepsilon) = 0, Cov(x_j, \varepsilon) = 0$
- 一致性只需要假设3’，但要满足无偏性需假设3
- ε 与自变量间存在相关性会使OLS没有一致性
- 一元回归的情形我们容易看到

$$p\lim \hat{\beta}_1 - \beta_1 = Cov(x_1, \varepsilon)/Var(x_1)$$

- 对缺失变量的问题去掉了变量 x_2 而进行回归得到 $\tilde{\beta}_1$ ，则有 $p\lim \tilde{\beta}_1 = \beta_1 + \beta_2\delta_1$

$$\delta_1 = Cov(x_1, x_2)/Var(x_1)$$

- 从实际问题出发，可以把不一致性和有偏同等看待，不一致性使用真实参数的方差和协方差，而有偏使用的是样本方差和协方差

39

3．3．2 渐近正态性和大样本推断

- OLS的正态性主要来源于假定了干扰项服从正态分布
- 如果 $\varepsilon_1, \cdots, \varepsilon_n$ 不是正态分布，则 $\hat{\beta}_j$ 也将不再是正态分布，接下来的t统计量也不再服从t分布，F统计量也不再服从F分布
- 因干扰项是不可观测的，而y是可以观测，可考虑y的分布是否象正态
- 即使在y不是正态分布时，仍然可以使用中心极限定理来得到OLS估计满足渐近正态性

40

- 定理3．3 （OLS的渐近正态性）在GM假设1~5之下，有：1）$\sqrt{n}(\hat{\beta}_j - \beta_j) \xrightarrow{a} N(0, \sigma^2/a_j^2)$
- 其中，$\sigma^2/a_j^2 > 0$是 $\sqrt{n}(\hat{\beta}_j - \beta_j)$ 的渐近方差，$a_j^2 = p\lim(n^{-1}\sum_{i=1}^n \hat{r}_{ij}^2)$ ，$\hat{r}_{ij}$ 为用其余自变量对 x_j 进行回归所得残差，称 $\hat{\beta}_j$ 渐近正态分布
- 2）$\hat{\sigma}^2 = (n-k-1)(n^{-1}\sum_{i=1}^n \hat{\varepsilon}_i^2)$是 $\sigma^2 = Var(\varepsilon)$ 的一致估计
- 3）对每一个j，$(\hat{\beta}_j - \beta_j)/se(\hat{\beta}_j) \xrightarrow{a} N(0,1)$ ，其中 $se(\hat{\beta}_j)$ 为通常OLS估计的标准差
- 结论3）实际上是t分 $(\hat{\beta}_j - \beta_j)/se(\hat{\beta}_j) \xrightarrow{a} t_{n-k-1}$

41

- OLS估计参数$\hat{\beta}_j$ 的方差为 $Var(\hat{\beta}_j) = \dfrac{\sigma^2}{SST_j(1-R_j^2)}$

$$SST_j(1-R_j^2) = SST_j \frac{(SST_j - SSE_j)}{SST_j} = SSR_j$$

- SSR_j为用其余自变量对 x_j 进行回归所得到的残差平方和 $SSR_j = \sum_{i=1}^n \hat{r}_{ij}^2$
- 结论1)中的 $a_j^2 = p\lim(n^{-1}SSR_j)$
- k=1时，SSR_1 即为 SST_1 ，此时有

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) = \sqrt{n}\frac{\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}$$

42

- 分母的极限是 σ_x^2
- 因为假设3，$(x_i-\bar{x})$ 与 ε_i 不相关，且是独立同分布的，它们之乘积$(x_i-\bar{x})\varepsilon_i$的期望为零，极限方差为 $\sigma_x^2\sigma^2$ 的，所以$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i-\bar{x})\varepsilon_i$ 渐近正态分布 $N(0,\sigma^2\sigma_x^2)$
- 因此 $\sqrt{n}(\hat{\beta}_j-\beta_j)$ 渐近于正态分布

$$N(0,\frac{\sigma^2\sigma_x^2}{(\sigma_x^2)^2})=N(0,\frac{\sigma^2}{\sigma_x^2})$$

- k>1时，对这一定理的证明比较复杂，不再详述

- 当样本量充分大时， t_{n-k-1} 趋向正态分布
- 没有合适的标准来判断到底多大的样本量才足够
- 样本量的大小还依赖于干扰项的分布
- 近似的好坏不仅依赖于n，而是n-k-1
- 定理需要同方差的假设
- OLS估计渐近正态的特性使得当样本量适当大时，F统计量也近似于F分布
- 相应的置信区间

- 3．3．3．拉格朗日乘子统计量
- 渐近分析框架下可有其它的方式来对多个约束条件进行假设检验
- LM统计量起源于有约束条件的优化
- 假定$\vec{\beta}_{ur}$是完全模型的极大似然估计，$\vec{\beta}_r$ 是限制模型的极大似然估计。则在约束条件 $\vec{\beta}_{ur}=\vec{\beta}_r$ 下，极大化对数似然函数，即极大化

$$\ln(L(\vec{\beta}_{ur}))-\lambda(\vec{\beta}_{ur}-\vec{\beta}_r)$$

- 完全模型和限制模型给出的参数估计比较接近，λ 的取值也会比较小；如果限制条件有显著的作用，加上这一约束条件的成本 λ 将会很大

- 有时也称这一检验方法为得分检验(score test)。LM检验的检验统计量为 $LM=\frac{[\lambda(\vec{\beta}_r)]^2}{I(\vec{\beta}_r)}$
- $\lambda(\vec{\beta}_r)$为对数似然的一阶偏导数，$I(\vec{\beta}_r)$为对数似然的二阶偏导数，也称为信息矩阵
- 导出LM统计量的形式只需要GM假设
- 考虑有k个自变量的回归模型

$$y=\beta_0+\beta_1x_1+\cdots+\beta_kx_k+\varepsilon$$

- 检验最后的q个自变量的真实参数都为零
- 零假设为 $H_0:\beta_{k-q+1}=0,\cdots,\beta_k=0$
- 备择假设也是其中至少有一个参数不是零

- LM统计量的一个优点是只需估计限制模型

$$y=\tilde{\beta}_0+\tilde{\beta}_1x_1+\cdots+\tilde{\beta}_{k-q}x_{k-q}+u$$

- 零假设下，u大致不应该与每一个自变量有关
- 用限制模型的回归残差通过其它自变量回归来得到LM检验　检验统计量为：LM=n R_u^2
- 可以验证在零假设下，样本量乘以附属回归的R方渐近于自由度为q的卡方分布
- q个变量约束的LM统计量的计算步骤为
- 1）用限制模型回归而得到残差序列u
- 2）用所有的自变量对u进行多元回归得到R方，记为：R_u^2

- 3）计算LM统计量LM=n R_u^2
- 4）比较LM与确定的服从 χ_q^2 分布的临界值c，如果LM>c则拒绝零假设。根据随机变量的分布计算p值
- 样本量比较大时，LM和F检验给出的结果基本没有差别，通常会使用F检验
- 例3．4 在CEO年薪的例子中，采用2007年的数据，只考虑公司的盈利和股东持股情况回归模型为：

$$\log(threeceo)=8.564+0.223\cdot\ln asset+0.356eachearn$$
$$-0.405firsthold+0.588\sec ondhold$$

- 残差平方和为：606.544
- 备择假设为：H_1 至少有一个 β_j 不为零，j=3,4

- 采用同样数据对限制模型进行估计得到

$$\log(threeceo) = 8.655 + 0.215 \cdot \ln asset + 0.339 eachearn$$

- 残差平方和为：**598.821**
- 计算出**F**统计量

$$F = [(606.544 - 598.821)/2]/(598.821/1050) \approx 6.76$$

- 将限制模型作为因变量，完全模型四个自变量代入数据得到回归模型为

$$\hat{usquare} = 0.004 \cdot \ln asset + 0.019 eachearn - 0.405 firsthold + 0.578 \sec ondhold$$

- 回归模型的**R**方为**0.00879**，计算**LM**统计量

$$\text{LM=n } R_u^2 = 0.00879 \times 1050 = 12.24$$

3．3．4．OLS估计的渐近有效性

- 定理3．4 （渐近有效性）在GM假设之下，假定 $\tilde{\beta}_j$ 为一阶约束条件 $\sum_{i=1}^{n} g_j(x_i)(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_{i1} - \cdots - \tilde{\beta}_k x_{ik}) = 0$ 的解，即 $\tilde{\beta}_j$ 为GLS估计。其中 $g_j(x)$ 为x的任意一个函数，则OLS估计在这类估计量具有最小的渐近方差 $AVar[\sqrt{n}(\hat{\beta}_j - \beta_j)] \le AVar[\sqrt{n}(\tilde{\beta}_j - \beta_j)]$
  - k=1时，回归模型为 $y = \beta_0 + \beta_1 x_1 + \varepsilon$
  - 假定 $g(x)$ 为x的任一函数，例如 $g(x) = x^2$
  - 则根据条件期望的性质，当 $E(\varepsilon \mid x) = E(\varepsilon) = 0$

- 可以推导出 $E(\varepsilon \mid g(x)) = E(\varepsilon) = 0$
- 定义一个新的变量为 $z_i = g(x_i)$
- 用类似OLS估计的方式得到一个新的参数估计为

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}$$

- 在g(x)与x是相关的条件下，得到的估计是一致的。将 $y = \beta_0 + \beta_1 x_1 + \varepsilon$ 代入的估计式得到

$$\begin{aligned}\tilde{\beta}_1 &= \frac{\sum_{i=1}^{n}(z_i - \bar{z})(\beta_0 + \beta_1 x_i + \varepsilon_i - (\beta_0 + \beta_1 \bar{x} + \bar{\varepsilon}))}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})} \\ &= \frac{\sum_{i=1}^{n}(z_i - \bar{z})(\beta_1(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon}))}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})} \\ &= \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(\varepsilon_i - \bar{\varepsilon})}{\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}\end{aligned}$$

- 根据大样本性质有

$$\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(\varepsilon_i - \bar{\varepsilon}) \xrightarrow{p} Cov(z, \varepsilon)$$

$$\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x}) \xrightarrow{p} Cov(z, x)$$

- 由假设3导出的条件保证了 $Cov(z, \varepsilon) = 0$
- 在 $Cov(z, x) \neq 0$ 时有 $p\lim \tilde{\beta}_1 = \beta_1$
- 此时有 $\sqrt{n}(\tilde{\beta}_1 - \beta_1) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(z_i - \bar{z})(\varepsilon_i - \bar{\varepsilon})}{\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}$
- 分母的极限是 $Cov(z, x)$
- $(z_i - \bar{z})$ 与 ε_i 不相关，且是独立同分布的
- 它们之乘积的期望为零，极限方差为 $Var(z)\sigma^2$
- $\sqrt{n}(\tilde{\beta}_1 - \beta_1)$ 渐近正态分布

$$\sqrt{n}(\tilde{\beta}_1 - \beta_1) \xrightarrow{p} N\left(0, \frac{\sigma^2 Var(z)}{(Cov(z, x))^2}\right)$$

- 用这一关系也可得到OLS估计的渐近分布，就是当z=x时，OLS估计的渐近方差为

$$\frac{\sigma^2 Var(x)}{(Cov(x, x))^2} = \frac{\sigma^2}{Var(x)}$$

- 据方差协方差的Cauchy-Schwartz不等式

$$[Cov(z, x)]^2 \le Var(z)Var(x)$$

- 可见 $\sqrt{n}(\tilde{\beta}_j - \beta_j)$ 的渐近方差不会小于OLS估计
- $\sqrt{n}(\hat{\beta}_j - \beta_j)$ 的渐近方差
- 对k>1的情形，有类似的结论

3．4．多元回归模型的一些特殊处理方法

- 3．4．1 回归模型的β 系数
- 模型系数不能告诉我们哪个变量的贡献更大
- 把所有的变量都作一个标准化
- 最初的回归模型为

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_{ik} + \hat{\varepsilon}_i$$

- 求平均再把它减回去

$$y_i - \bar{y} = \hat{\beta}_1(x_{i1} - \bar{x}_1) + \cdots + \hat{\beta}_k(x_{ik} - \bar{x}_k) + \hat{\varepsilon}_i$$

- 因变量的样本标准差调整方程为

$$(y_i - \bar{y})/\hat{\sigma}_y = (\hat{\sigma}_1/\hat{\sigma}_y)\hat{\beta}_1[(x_{i1} - \bar{x}_1)/\hat{\sigma}_1] + \cdots$$

$$+ (\hat{\sigma}_k/\hat{\sigma}_y)\hat{\beta}_k[(x_{ik} - \bar{x}_k)/\hat{\sigma}_k] + (\hat{\varepsilon}_i/\hat{\sigma}_y)$$

- 自变量的系数正好是该变量原来的系数乘以自变量的标准差再除以因变量y的标准差
- 作一个变量替换

$$\hat{b}_j = (\hat{\sigma}_j/\hat{\sigma}_y)\hat{\beta}_j \qquad z_{ij} = (x_{ij} - \bar{x}_j)/\hat{\sigma}_j$$

$$(y_i - \bar{y})/\hat{\sigma}_y = \hat{b}_1 z_{i1} + \cdots + \hat{b}_k z_{ik} + \hat{u}_i$$

- 新系数 $\hat{b}_j$ 通常称为标准化系数或贝塔系数
- 自变量增加一个标准差，y改变了 $\hat{b}_j$ 个标准差
- 标准化模型中，贝塔系数就是可比较的

3．4．2 模型中的平方项

- 表示边际影响有逐渐增加或减少趋势
- 工龄的增加对工资增长的影响可能是逐渐减少

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

$$\Delta\hat{y} \approx (\hat{\beta}_1 + 2\hat{\beta}_2 x)\Delta x$$

- 表示x与y之间的关系与x的水平有关
- 代入数据 $\widehat{wage} = 3.73 + 0.298\exp er - 0.0061\exp er^2$

（0.35） （0.041） (0.0009)

随x水平的增加对y的影响由正转负的转折点

$x^* = |\hat{\beta}_1/2(\hat{\beta}_2)|$ $x^* = |0.298/(2\times 0.0061)| \approx 24.4$

3．4．3 模型中的交叉项

- 两个变量有交叉效应
- 适当的参数重组来表达我们所关注的变量

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

- 只简单地说当 x_1 =1时，x_2 对y的偏效应为 β_3，显然不能表达这一模型
- 把模型进行调整

$$y = \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \beta_3(x_1 - u_1)(x_2 - u_2) + \varepsilon$$

$$\delta_1 = \beta_1 + \beta_3 u_2 \qquad \delta_2 = \beta_2 + \beta_3 u_1$$

- 在 x_1 的均值附近，x_2 对y的偏效应是 $\delta_2 = \beta_2 + \beta_3 u_1$

3．4．4 对模型拟合度的调整

- R方衡量因变量y的方差有多少被自变量所解释
- 干扰项的方差大不容易得到准确的估计
- 样本量的增加能够抵消误差项方差的影响
- 加入变量R方的相对改变量
- 调整的R方：把R方定义的分子分母同除n

$$R^2 = 1 - (SSR/n)/(SST/n)$$

- 分别成为因变量和干扰项方差的估计
- 模型真实的R方为 $1 - \sigma_\varepsilon^2/\sigma_y^2$

- 估计是有偏的
- 如果使用无偏估计，我们就得到了调整的R方

$$\bar{R}^2 = 1 - [SSR/(n-k-1)]/[SST/(n-1)]$$

- 调整R方对增加自变量给出了惩罚因子
- 增加一个自变量就不一定会导致它增加
- 调整R方增大当且仅当新变量的t统计量大于1
- 增加一组自变量，调整R方增大当且仅当这一组变量的F统计量大于1
- 例 R^2=0.30，n=51，k=10,$\bar{R}^2$=1-0.7×50/40=0.125
- 调整R方可能为负

3．4．5使用调整R方对非嵌套模型进行筛选

- F检验只能对两个存在包含关系的模型进行检验
- 采用调整R方进行比较就可以直接判断两个模型的优劣
- 考虑公司的R&D投入对销售收入的影响

$$rdi = \beta_0 + \beta_1 \log(sales) + \varepsilon$$

$$rdi = \beta_0 + \beta_1 sales + \beta_2 sales^2 + \varepsilon$$

- 直接使用R方前一模型为0.061后一模型为0.148
- 使用调整R方前一模型为0.030后一模型为0.090

3．4．6 回归模型多余变量的控制

- 模型中包含的变量过多
- 应用中到底哪些变量需要控制，哪些不需要控制通常并不那么明确
- 如能明确知道它与自变量无关，加入模型不会带来多重共线性问题，还能提高参数估计精度
- 若数据可得到，又不会产生多重共线性问题，通常还是把它包含在模型中

3．4．7使用回归模型进行预测

- 用置信区间预测
- 假定 $c_1,\cdots c_k$ 是k个自变量的一组取值则有

$$\hat{\theta}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \cdots + \hat{\beta}_k c_k$$

- 以 $\hat{\theta}_0$ 为中心来构造置信区间
- 需要得到OLS估计的一个线性组合的标准差记

$$\beta_0 = \theta_0 - \beta_1 c_1 - \cdots - \beta_k c_k$$

$$y = \theta_0 + \beta_1 (x_1 - c_1) + \beta_2 (x_2 - c_2) + \cdots + \beta_k (x_k - c_k) + \varepsilon$$

- 这一回归模型截距项的标准差就是我们所求
- 当取值在自变量均值附近，截距项标准差比较小

- 没有自变量观测值对应的置信区间需要考虑不可观测的影响，预测区间
- y^0 可能代表某个不在样本内的个人或公司
- $x_1^0,\cdots,x_k^0$ 为假想能观测的对应的自变量取值
- ε^0为不可观测的误差

$$y^0 = \beta_0 + \beta_1 x_1^0 + \cdots + \beta_k x_k^0 + \varepsilon^0$$

- 最优期望值为 $\hat{y}^0 = \hat{\beta}_0 + \hat{\beta}_1 x_1^0 + \cdots + \hat{\beta}_k x_k^0$
- 预测误差的方差为 $Var(\hat{e}^0) = Var(\hat{y}^0) + Var(\varepsilon^0)$
- $Var(\hat{y}^0)$是因为$\hat{\beta}_j$为估计量而带来的误差
- $Var(\hat{y}^0)$ 的方差是1/n的比例

总结

- 利用最小二乘估计的分布特性对模型参数的偏效应是否显著进行统计推断的t检验和F检验
- 在干扰项不满足正态分布的时候，OLS估计的渐近性：估计的一致性
- OLS估计的渐近正态性和大样本时使用LM统计量，给出了OLS渐近有效的范围
- 模型中的一些特殊处理方法：标准化的模型系数，平方项和交叉项在模型中的作用
- 调整R方及如何使用模型进行预测
- 通过残差分析来考察模型和特殊数据点。

再见！