线性回归-运用篇

李思芃

2020.10.29

本章框架

- 3.4：介绍样本最小二乘法及其估计系数性质
- 3.5：有限样本和大样本假设检验
- 3.6：回归方法Stata命令实例
- 3.7：回归分析运用中的常见问题

3.4 样本最小二乘法及其估计系数性质

- 实际研究中，用样本估计总体的线性关系系数$\hat{\beta}$
- 拟合值 $\widehat{Y_i} = \boldsymbol{X_i'}\widehat{\boldsymbol{\beta}}$ 与观测值之差 $\hat{e}_i = Y_i - \hat{Y}_i$ 的平方和均值最小

$$\widehat{\boldsymbol{\beta}} = \underset{b}{argmin}\frac{1}{N}\sum_i^N (Y_i - \boldsymbol{X_i'}b)^2$$

由对b的一阶导数等于0可得

$$\sum_i^N \boldsymbol{X_i}\left(Y_i - \boldsymbol{X_i'}\widehat{\boldsymbol{\beta}}\right) = 0$$

解出$\widehat{\boldsymbol{\beta}}$为

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{\mathrm{i}=1}^N \boldsymbol{X_i}\boldsymbol{X_i'}\right)^{-1}\sum_{i=1}^N \boldsymbol{X_i}Y_i$$

3.4 样本最小二乘法及其估计系数性质

用矩阵表示，

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

- 将$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$代入可得样本估计值与真实值的关系：

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X\beta} + \boldsymbol{e}) = \boldsymbol{\beta} + \underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{e}}_{\text{误差项}}$$

误差项具体展开：

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{e} = (\sum_{i=1}^{N}\boldsymbol{X}_i\boldsymbol{X}_i')^{-1}\sum_{i=1}^{N}\boldsymbol{X}_i\, e_i =$$

$$(\sum_{i=1}^{N}\boldsymbol{X}_i\boldsymbol{X}_i')^{-1}\begin{bmatrix}\sum_{i=1}^{N}\boldsymbol{X}_{1i}\, e_i\\ \sum_{i=1}^{N}\boldsymbol{X}_{2i}\, e_i\\ \vdots\\ \sum_{i=1}^{N}\boldsymbol{X}_{Ni}\, e_i\end{bmatrix}$$

3.4 样本最小二乘法及其估计系数性质

- 固定解释变量（3.4.1）、随机解释变量（3.4.2）
- 有限样本性质　无偏性、方差、分布
- 大样本性质　　一致性、分布

3.4.1 固定解释变量下的样本估计系数性质

- 有限样本性质
  - 无偏性

$$E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E[\boldsymbol{e}] = \beta$$

  - 方差

$$Var[\hat{\boldsymbol{\beta}}] = \mathrm{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] = \mathrm{E}[((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{e})((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{e})']$$

$$= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{E}[\boldsymbol{e}\boldsymbol{e}']\boldsymbol{X}\,(\boldsymbol{X}'\boldsymbol{X})^{-1} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\sigma}^{2}\boldsymbol{I}\boldsymbol{X}\,(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

$$= \sigma^{2}\,(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

干扰项方差σ^2越大，估计值系数分布的分散度越大，偏离真实值的范围越大

3.4.1 固定解释变量下的样本估计系数性质

- 分布

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{e}$$

$$\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta},\ \sigma^2\ (\boldsymbol{X}'\boldsymbol{X})^{-1})$$

$$\hat{\beta}_{\mathrm{k}} \sim N(\beta_k\ ,\ \sigma^2\ (\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk})$$

- 大样本性质

没有对$e \sim N(0, \boldsymbol{\sigma^2 I})$的假设，但样本量足够大的时候，估计系数仍能达到理想的性质

- 一致性

$$plim\ \widehat{\boldsymbol{\beta}} = plim(\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{e})$$

$$= \boldsymbol{\beta} + lim\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{N}\right)^{-1} plim(\frac{\boldsymbol{X}'\boldsymbol{e}}{N})$$

$$= \boldsymbol{\beta} + \boldsymbol{Q}^{-1}\ plim(\frac{\boldsymbol{X}'\boldsymbol{e}}{N})$$

3.4.1 固定解释变量下的样本估计系数性质

- 一致性

判断$plim(\frac{\boldsymbol{X'e}}{N}) = 0$

运用均值平方收敛，即当样本数量N趋近于无穷，若$E(\frac{\boldsymbol{X'e}}{N}) = 0, Var(\frac{\boldsymbol{X'e}}{N}) = 0$, 则$\frac{\boldsymbol{X'e}}{N}$均值平方收敛于0。若一个随机变量均值平方收敛与某个常数，则它必然概率收敛于该常数

$$E(\frac{\boldsymbol{X'e}}{N}) = \frac{1}{N}E(\sum_{i=1}^{N}\boldsymbol{X_i}e_i) = \frac{1}{N}\sum_{i=1}^{N}E(\boldsymbol{X_i}e_i) = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{X_i}E(e_i) = 0$$

$$Var(\frac{\boldsymbol{X'e}}{N}) = \frac{1}{N^2}Var(\sum_{i=1}^{N}\boldsymbol{X_i}e_i) = \frac{\sum_{i=1}^{N}X_iX_i'}{N}\frac{\sigma^2}{N} = 0$$

3.4.1 固定解释变量下的样本估计系数性质

- 分布

根据Lindeberg-Levy中心极限定理

$$\frac{1}{\sqrt{N}}X'\mathrm{e} \xrightarrow{\mathrm{d}} N[0,\sigma^2\boldsymbol{Q}]$$

$$\boldsymbol{Q}^{-1}\frac{1}{\sqrt{N}}X'\mathrm{e} \xrightarrow{\mathrm{d}} N[0,\boldsymbol{Q}^{-1}(\sigma^2\boldsymbol{Q})\boldsymbol{Q}^{-1}]$$

$$\sqrt{N}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}) \xrightarrow{\mathrm{d}} N[0,\sigma^2\boldsymbol{Q}^{-1}]$$

3.4.2 随机解释变量下的样本估计系数性质

- 小样本性质
  - 无偏性

$$E(\hat{\boldsymbol{\beta}}) = E_X\left[E\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right)\right] = E_X[\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{e} \mid \boldsymbol{X})] = E_X(\boldsymbol{\beta}) = \boldsymbol{\beta}$$

假设：$E(\boldsymbol{e}|\boldsymbol{X}) = 0$

当$\boldsymbol{X}$为一个固定值x的时候，$E\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X} = \boldsymbol{x}\right) = \boldsymbol{\beta}$

对所有$\boldsymbol{X}$的可能只进行重复实验，最后把所有实验的估计系数平均值按$\boldsymbol{X}$的分布概率取期望，则$E_X\left[E\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right)\right] = E_X(\boldsymbol{\beta}) = \boldsymbol{\beta}$

  - 方差

$$Var\left[\hat{\boldsymbol{\beta}}\right] = \mathrm{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] = E_X[\mathrm{E}((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mid \boldsymbol{X})]$$

$$= E_X[\,(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E[\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{X}]\boldsymbol{X}\,(\boldsymbol{X}'\boldsymbol{X})^{-1}] = \sigma^2\,E(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

3.4.2 随机解释变量下的样本估计系数性质

- 分布

$$e \mid \boldsymbol{X} \sim \mathrm{N}(0,\sigma^2)$$

$$\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}\boldsymbol{X}')^{-1})$$

$$\hat{\beta}_{\mathrm{k}} \mid X \sim N(\beta_k\,,\ \sigma^2\ (\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk})$$

- 大样本性质
  - 一致性

$$plim\ \widehat{\boldsymbol{\beta}} = plim(\boldsymbol{\beta}+(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{e})$$

$$= \boldsymbol{\beta} + \mathrm{p}lim\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{N}\right)^{-1} plim\left(\frac{\boldsymbol{X}'\boldsymbol{e}}{N}\right)$$

$$= \boldsymbol{\beta} + \boldsymbol{Q}^{-1}\ plim(\frac{\boldsymbol{X}'\boldsymbol{e}}{N})$$

3.4.2 随机解释变量下的样本估计系数性质

- 分布

$$\sqrt{N}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}) \xrightarrow{\mathrm{d}} N[0,\sigma^2\boldsymbol{Q}^{-1}]$$

3.5 有限样本和大样本假设检验

- 样本数量大小会对估计系数的分布产生影响
- 有限样本假设检验
  - 固定解释变量

若干扰项e服从正态分布，则$\hat{\beta}_{\mathrm{k}} \sim N(\beta_k, \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk})$

若σ^2 已知，使用Z统计量检验：

$$Z = \frac{\hat{\beta}_{\mathrm{k}} - \beta_k}{\sqrt{V_{\mathrm{k}}}} \sim \mathrm{N}(0,1)$$

若σ^2 未知，σ^2的一个无偏估计 $s^2 = \frac{\hat{e}'\hat{e}}{n-K}$，$\hat{V}_k = [s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{kk}$

使用t统计量检验：

$$t = \frac{\hat{\beta}_{\mathrm{k}} - \beta_k}{\sqrt{\hat{V}_k}}$$

3.5 有限样本和大样本假设检验

- 随机变量

$$t \mid \boldsymbol{X} = \frac{\hat{\beta}_{\mathrm{k}} - \beta_k}{\sqrt{\hat{V}_k}}$$

该统计量的无条件分布和有条件分布都是自由度等于$N-k$的t分布

3.5 有限样本和大样本假设检验

- 大样本假设检验（如果干扰项不服从正太分布）
  - 固定解释变量（随机变量）

$$\sqrt{N}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}) \xrightarrow{\mathrm{d}} N[0,\sigma^2\boldsymbol{Q}^{-1}]$$

$$\sqrt{N}(\beta-\hat{\beta}_{\mathrm{k}}) \xrightarrow{\mathrm{d}} N[0,\ \sigma^2\boldsymbol{Q}_{kk}^{-1}]$$

用无偏估计 $s^2=\frac{\hat{e}'\hat{e}}{n-K}$ 估计σ^2，则估计系数的大样本方差一致估计量为：

$$\widehat{Avar(\hat{\beta}_k)}=[s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{kk}$$

用Z统计量进行假设检验：

$$Z=\frac{\hat{\beta}_{\mathrm{k}}-\beta_k}{\widehat{Avar(\hat{\beta}_k)}}\sim \mathrm{N}(0,1)$$

3.6 回归方法Stata命令实例

```
. use "C:\Users\lsp\Desktop\STATA数据.dta"

. describe

Contains data from C:\Users\lsp\Desktop\STATA数据.dta
  obs:           168
 vars:             7                          24 Oct 2020 15:30
 size:         2,184
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
sex             int     %8.0g                 male=1,female=0
African         int     %8.0g                 African=1,not African=0
Hispanic        int     %8.0g                 Hispanic=1,not Hispanic=0
Asian           int     %8.0g                 Asian=1,not Asian=0
middleses       int     %8.0g                 middleclass=1,not middleclass=0
highses         int     %8.0g                 highclass=1,not highclass=0
read            byte    %10.0g                read
-------------------------------------------------------------------------------
```

3.6 回归方法Stata命令实例

```
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
sex	168	.4642857	.5002138	0	1
African	168	.1071429	.3102194	0	1
Hispanic	168	.1190476	.3248127	0	1
Asian	168	.047619	.2135955	0	1
middleses	168	.4464286	.498608	0	1
highses	168	.2738095	.4472455	0	1
read	168	51.84524	10.42279	28	76

```
. summarize read, detail
```

read

Percentiles		Smallest		
1%	31	28		
5%	35	31		
10%	39	34	Obs	168
25%	44	34	Sum of Wgt.	168
50%	50		Mean	51.84524
		Largest	Std. Dev.	10.42279
75%	60	73		
90%	66	73	Variance	108.6346
95%	68	76	Skewness	.2038318
99%	76	76	Kurtosis	2.350443

3.6 回归方法Stata命令实例

```
. tabstat sex African Hispanic Asian middleses highses read, stats(n mean sd min max p25 p50 p75) format(%9.0g) column(sta
> tistics)
```

variable	N	mean	sd	min	max	p25	p50	p75
sex	168	.4642857	.5002138	0	1	0	0	1
African	168	.1071429	.3102194	0	1	0	0	0
Hispanic	168	.1190476	.3248127	0	1	0	0	0
Asian	168	.047619	.2135955	0	1	0	0	0
middleses	168	.4464286	.498608	0	1	0	0	1
highses	168	.2738095	.4472455	0	1	0	0	1
read	168	51.84524	10.42279	28	76	44	50	60

3.6 回归方法Stata命令实例

- 回归前执行correlate考察变量相关性

```
. correlate
(obs=168)
```

	sex	African	Hispanic	Asian	middle~s	highses	read
sex	1.0000						
African	-0.0524	1.0000					
Hispanic	0.0632	-0.1273	1.0000				
Asian	-0.1521	-0.0775	-0.0822	1.0000			
middleses	0.0523	-0.1175	0.0396	-0.0321	1.0000		
highses	0.0707	-0.0832	-0.1433	0.0507	-0.5514	1.0000	
read	0.0633	-0.1782	-0.2351	-0.0101	-0.0707	0.2905	1.0000

3.6 回归方法Stata命令实例

```
. regress read sex African Hispanic Asian middleses highses
```

Source	SS	df	MS	Number of obs	=	168
				F(6, 161)	=	5.28
Model	2983.77159	6	497.295265	Prob > F	=	0.0001
Residual	15158.2046	161	94.1503391	R-squared	=	0.1645
				Adj R-squared	=	0.1333
Total	18141.9762	167	108.634588	Root MSE	=	9.7031

read	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sex	.7784912	1.537179	0.51	0.613	-2.257143	3.814125
African	-5.959343	2.517923	-2.37	0.019	-10.93176	-.9869286
Hispanic	-7.26791	2.380875	-3.05	0.003	-11.96968	-2.566138
Asian	-2.383286	3.586637	-0.66	0.507	-9.466205	4.699634
middleses	1.43899	1.857729	0.77	0.440	-2.229668	5.107648
highses	6.549484	2.093515	3.13	0.002	2.415195	10.68377
_cons	50.6653	1.701195	29.78	0.000	47.30576	54.02483

3.6 回归方法Stata命令实例

- 检验两个系数是否相等

```
. test Asian=African

 ( 1)  - African + Asian = 0

       F(  1,   161) =    0.73
            Prob > F =    0.3944

. test highses middleses

 ( 1)  highses = 0
 ( 2)  middleses = 0

       F(  2,   161) =    5.68
            Prob > F =    0.0041
```

```
. test sex highses African

 ( 1)  sex = 0
 ( 2)  highses = 0
 ( 3)  African = 0

       F(  3,   161) =    6.71
            Prob > F =    0.0003
```

- 检验变量是否有共同显著的作用

3.6 回归方法Stata命令实例

- 使用不同模型估计时，可以用estimate store把不同模型回归结果储存起来，然后再用esttab命令生成表格

```
. reg read African Hispanic Asian middleses highses

. estimate store reg1, title(regression 1)

. reg sex African Hispanic Asian

. estimate store reg2, title(regression 2)

. esttab reg1 reg2, b(%7.3f) se(%7.3f) stat(N r2 F) title("回归结果")
```

3.6 回归方法Stata命令实例

回归结果

	(1) read	(2) read
African	-5.995*	-7.129**
	(2.511)	(2.524)
Asian	-2.662	-1.907
	(3.536)	(3.685)
Hispanic	-7.191**	-8.642***
	(2.371)	(2.409)
middleses	1.538	
	(1.843)	
highses	6.685**	
	(2.072)	
sex		1.317
		(1.567)
_cons	50.953***	53.117***
	(1.600)	(1.179)
N	168.000	168.000
r2	0.163	0.106
F	6.316	4.807

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

3.7 回归分析运用中的常见问题

- 对线性回归系数的解释
- 对变量取对数
- 在线性模型中加入二次项
- 缩放变量
- 移动变量
- 统计显著性与经济显著性
- 包含交叉项 的模型
- 多余的解释变量
- 多重共线性
- 检验分组系数的不同

3.7.1 对线性回归方程的解释

- 以两变量线性回归为例：

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + e, E(e \mid X_1, X_2) = 0$$

  - 对总体系数β_1解释

$$E(Y \mid X_1 = x_1, X_2 = x_2) - E(Y \mid X_1 = x_1 + 1, X_2 = x_2) = \beta_1$$

  - 对样本估计系数的解释

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

$$\hat{Y}_1 = \hat{\alpha} + \hat{\beta}_1 (X_1 + 1) + \hat{\beta}_2 X_2$$

$$\hat{\beta}_1 = \hat{Y} - \hat{Y}_1$$

3.7.2 对变量取对数

- 变量关系非线性（画散点图）
- 取对数意味着原解释变量对解释变量的弹性，即百分比变化而非数值变化
- 对解释变量和被解释变量是否取对数有3种情况：
  - 对被解释变量取自然对数：

$$\widehat{\ln(Y)} = \hat{\alpha} + \hat{\beta}X$$

$$\widehat{\ln(Y')} = \hat{\alpha} + \hat{\beta}(X+1)$$

$$\widehat{\ln(Y')} - \widehat{\ln(Y)} = \ln\left(\frac{\widehat{Y'}}{\hat{Y}}\right) = \ln\left(1 + \frac{\Delta\hat{Y}}{\hat{Y}}\right) = \hat{\beta}$$

3.7.2 对变量取对数

- 对被解释变量取自然对数

若使用自然对数近似的性质：当m很小时，$\ln(1+\mathrm{m})\approx \mathrm{m}$

故有 $\frac{\Delta\hat{Y}}{\hat{Y}}=\hat{\beta}$

精确关系：$\frac{\Delta\hat{Y}}{\hat{Y}}=\mathrm{e}^{\hat{\beta}}-1$

- 对解释变量取自然对数

$$\hat{Y}=\hat{\alpha}+\hat{\beta}\ln X$$

$$\Delta\hat{Y}=\hat{\beta}(\ln X'-\ln X)=\hat{\beta}\ln(1+\frac{\Delta X}{X})$$

近似关系：$\Delta\hat{Y}=\hat{\beta}\frac{\Delta X}{X}$

精确关系：$\Delta\hat{Y}=\hat{\beta}\ln(1+\frac{\Delta X}{X})$

3.7.2 对变量取对数

- 对解释变量和被解释变量同时取自然对数

$$\widehat{\ln(Y)} = \hat{\alpha} + \hat{\beta} \ln X$$

$$\widehat{\ln(Y')} - \widehat{\ln(Y)} = \hat{\beta} (\ln X' - \ln X)$$

$$\ln\left(1 + \frac{\Delta \hat{Y}}{\hat{Y}}\right) = \hat{\beta} \ln\left(1 + \frac{\Delta X}{X}\right)$$

近似关系：$\frac{\Delta \hat{Y}}{\hat{Y}} = \hat{\beta} \frac{\Delta X}{X}$

精确关系：$\frac{\Delta \hat{Y}}{\hat{Y}} = e^{\hat{\beta} \ln\left(1 + \frac{\Delta X}{X}\right)} - 1$

3.7.3 在线性模型中加入二次项

- 二次方程

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$$

通过求导，可以发现X对Y的作用取决于X的值

变形：

$$\hat{Y} = \hat{\beta}_2 (X + \frac{\hat{\beta}_1}{2\hat{\beta}_2})^2 + (\hat{\beta}_0 - \frac{\hat{\beta}^{1^2}}{4\hat{\beta}_2})$$

对称轴： $X = -\frac{\hat{\beta}_1}{2\hat{\beta}_2}$

$\hat{\beta}_1 < 0, \hat{\beta}_2 > 0$

注意X的取值范围

3.7.4 缩放变量

- 模型$\hat{Y} = \hat{\alpha} + \hat{\beta}X$
  - 被解释变量变化n倍

$$n\hat{Y} = n\hat{\alpha} + n\hat{\beta}X \Rightarrow \hat{Y}^* = \hat{\alpha}^* + \hat{\beta}^*X$$

  - 将解释变量变化m倍

$$\hat{Y} = \hat{\alpha} + \frac{\hat{\beta}}{m}mX \Rightarrow \hat{Y}^* = \hat{\alpha} + \hat{\beta}^*X^*$$

  - 将被解释变量变化n倍，解释变量变化m倍

$$n\hat{Y} = n\hat{\alpha} + n\frac{\hat{\beta}}{m}mX \Rightarrow \hat{Y}^* = \hat{\alpha}^* + \hat{\beta}^*X^*$$

方程系数经济含义和系统显著性不变,只是单位发生改变，更易解释

3.7.4 缩放变量

- 对数形式缩放

$$\widehat{\ln(Y)} = \hat{\alpha} + \hat{\beta}\ln X$$

$$\widehat{\ln(nY)} - \ln n = \hat{\alpha} + \hat{\beta}\ln mX - \hat{\beta}\ln m$$

$$\widehat{\ln(Y)} = \hat{\alpha} + \hat{\beta}\ln X$$

$$\widehat{\ln(Y^*)} = (\hat{\alpha} + \ln n - \hat{\beta}\ln m) + \hat{\beta}\ln X^*$$

可以看出，只有截距改变，斜率不变，因为对数形式下系数反映的是百分比影响

3.7.5 移动变量

- 通过给变量增加或减去一定的值来改变变量大小
- $\hat{Y} = \hat{\alpha} + \hat{\beta} X$
- 被解释变量$+ q$：$\hat{Y} + q = \hat{\alpha} + q + \hat{\beta} X$
- 解释变量+p：$\hat{Y} + q + \hat{\beta} p = \hat{\alpha} + q + \hat{\beta}(X + p)$
- 得到：$\hat{Y}^* = (\hat{\alpha} + q - \hat{\beta} p) + \hat{\beta} X^*$

3.7.6 统计显著性和经济显著性（实际显著性）

- 统计显著性：对估计参数值进行假设检验来检查是否显著不同于原假设假定的值。通常在原假设中假设系数真值为0，然后计算偶然事件发生的概率p值，若p值小于显著性水平，就拒绝原假设，认为系数是起作用的
- 统计显著不意味着实际显著“统计显著性”主要是为了分析在小样本 中,那些看起来具有“实际显著性”的现象是否只是 偶然发生的。相反地,当样本数量很大时,虽然回归 系数也许能通过显著性检验,但它带来的效应可能 小到并不具有实际意义
- 统计显著性与经济显著性的六种组合

3.7.6 统计显著性和经济显著性（实际显著性）

- 不一致的成因
  - 抽样数据所得到的各项参数估计包含的抽样误差

统计显著与否受三个因素影响：实际效应强度的大小 、置信度水平的大小、样本规模的大小

  - 回归计算建立在严格的假设上，比如随机扰动项服从正态分布，拥有相同的均值和方差。但实际中由于未观测的因素，随机干扰项的方差几乎不可能相同，还可能相关
  - 假设检验基于小概率原理，μ=μ0 和检验统计量值对应的 p 值小于 0.05 有极大的概率相联系,但它们的关系是仍然是随机的,它们之间不存在演绎推论关系。这种显著是在一定较小允许误差下显著，差异不显著可能只是因为偶然因素、误差过大导致，并不能排除今后出现差异显著的可能性

《统计显著性与实际显著性辨析》

3.7.7 包含交叉项的模型

- 考虑模型

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + u$$

  - 一定要包含主项，否则由于确实变量偏差，交互作用可能是错误的
  - 解释系数：$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_1 X_2$

$\hat{\beta}_1(X_2 = 0)$、$\hat{\beta}_2$

$\hat{\beta}_3$：交叉项系数反映一个变量对$\hat{Y}$的影响是否受到另一个变量大小的影响

判断正误：“当$\hat{\beta}_1 < 0, \hat{\beta}_3 > 0$时，$X_1$的增加会使得$\hat{Y}$值增加，但$X_1$对$\hat{Y}$的影响会随着$X_2$的增加而减少”

无法通过该条件判断X_1对$\hat{Y}$的影响，取决于X_2

3.7.8 多余的解释变量

- 假设模型的真实情况：

$$Y = \beta_0 + \beta_1 X_1 + e$$

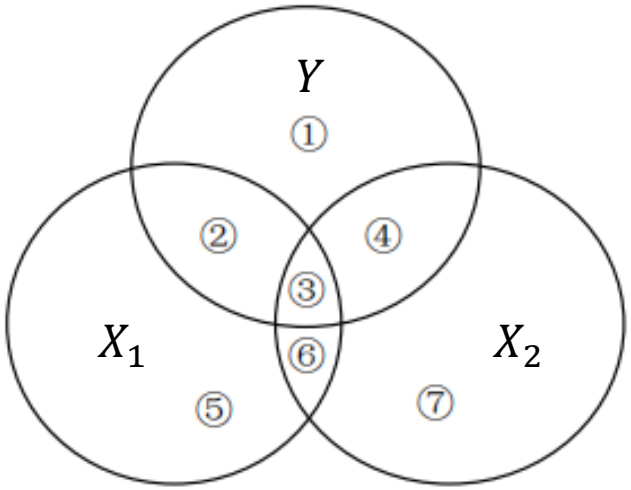
- 加入多余解释变量：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$$

  - $E(\hat{\beta}_1) = \beta_1, E(\hat{\beta}_2) = 0$
  - $Var(\hat{\beta}_j) = \dfrac{\sigma^2}{\sum_i^N (X_{ji} - \bar{X}_j)^2 (1 - R_j^2)}$
  - $(X_{ji} - \bar{X}_j)^2$、σ^2、R_j^2（共线性）

3.7.8 多余的解释变量

- 多余自变量对被解释变量没有解释力度，不能降低σ^2，且若多余自变量与解释变量j相关，会造成R_j^2变大，从而导致系数方差变大



3.7.9 多重共线性

- 考虑模型

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e$$

若X_2、X_3高度相关，X_1与其他解释变量无关

R_1^2=0，$Var(\hat{\beta}_1)$不受X_2、X_3高度相关的影响

但R_2^2、$R_3^2 > 0$，$Var(\hat{\beta}_{2,3})$变大，易被判断为不显著

3.7.10 检验分组系数的不同

- 样本分组，检验回归系数在不同组是否显著不同
- 考虑财务杠杆对企业投资影响的模型

$$Investment_i = \alpha + \beta_1 leverage_i + \beta_2 Size_i + \beta_3 Q_i + e_i$$

（Q是托宾Q值，企业业绩增长衡量指标）

- 假设β_1显著为负，进一步检验财务杠杆对投资量的影响是否对于有高增长机会的企业更严重
- 如何分组？如果Q大于样本中值，则为高增长企业；Q小于样本中值，为低增长企业

3.7.10 检验分组系数的不同

- 全样本全交乘项

$$Investment_i = \alpha + \beta_{11} leverage_i + \beta_{12} leverage_i \times HighQ_i + \beta_{21} Size_i + \beta_{22} Size_i \times HighQ_i + \beta_3 Q_i + e_i$$

β_{11}、 β_{12}、 $\beta_{11} + \beta_{12}$

好处：使用全样本数据、允许其他控制变量不同组的系数是不同的

缺点：如果控制变量数量较多，需要加入许多交乘项

3.7.10 检验分组系数的不同

- 全样本单交乘项

$$Investment_i = \alpha + \beta_{11} leverage_i + \beta_{12} leverage_i \times HighQ_i + \beta_2 Size_i + \beta_3 Q_i + e_i$$

β_{11}、 β_{12}

避免许多交乘项

缺点：隐含地限制了$Size$在不同组对投资的影响都相同

$$Investment_i = \alpha + \beta'_{11} leverage_i \times LowQ_i + \beta'_{12} leverage_i \times HighQ_i + \beta_2 Size_i + \beta_3 Q_i + e_i$$

使用两组虚拟变量： $HighQ_i$、 $LowQ_i$

$\beta'_{11} = \beta_{11}$， $\beta'_{12} = \beta_{12} + \beta_{11}$

通过$\beta'_{12} - \beta'_{11}$是否显著不等于0检验$leverage$在两组是否有显著差异

3.7.10 检验分组系数的不同

- 分样本回归

将样本按Q分成两组，分别进行回归

Q大于样本均值：

$$Investment_i = \alpha + \beta_{11} leverage_i + \beta_{21} Size_i + \beta_3 Q_i + e_i$$

Q小于样本均值：

$$Investment_i = \alpha + \beta_{12} leverage_i + \beta_{12} Size_i + \beta_3 Q_i + e_i$$

优点：允许不同组系数不同

缺点：检验$\beta_{12} - \beta_{11}$的显著性