Multiple Regression Analysis：Estimation

1. 一点提示：

- 统计学：SSR（regression）SSE（error），计量经济学：SSE（explained）SSR（residual）
- 自由度：
  - 伍：$F=\frac{SSE/k}{SSR/(n-k-1)}$
  - 张,南：$F=\frac{RSS/k}{ESS/(n-k-1)}$
  - 张,清：$F=\frac{ESS/(k-1)}{RSS/(T-k)}$
  - 伍：$F=\frac{(SSR_r-SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$
  - 陈：$F=\frac{(SSR^*-SSR)/m}{SSR/(n-K)}$

- SR与MR的比较（相同、不同与区别转化）：

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

$$Var(\widehat{\beta}_1)_{mr} = Var(\widehat{\beta}_1)_{sr} \times VIF$$

- 回归结果解读：

$$t = \frac{\widehat{\beta}_i}{se(\widehat{\beta}_i)}$$

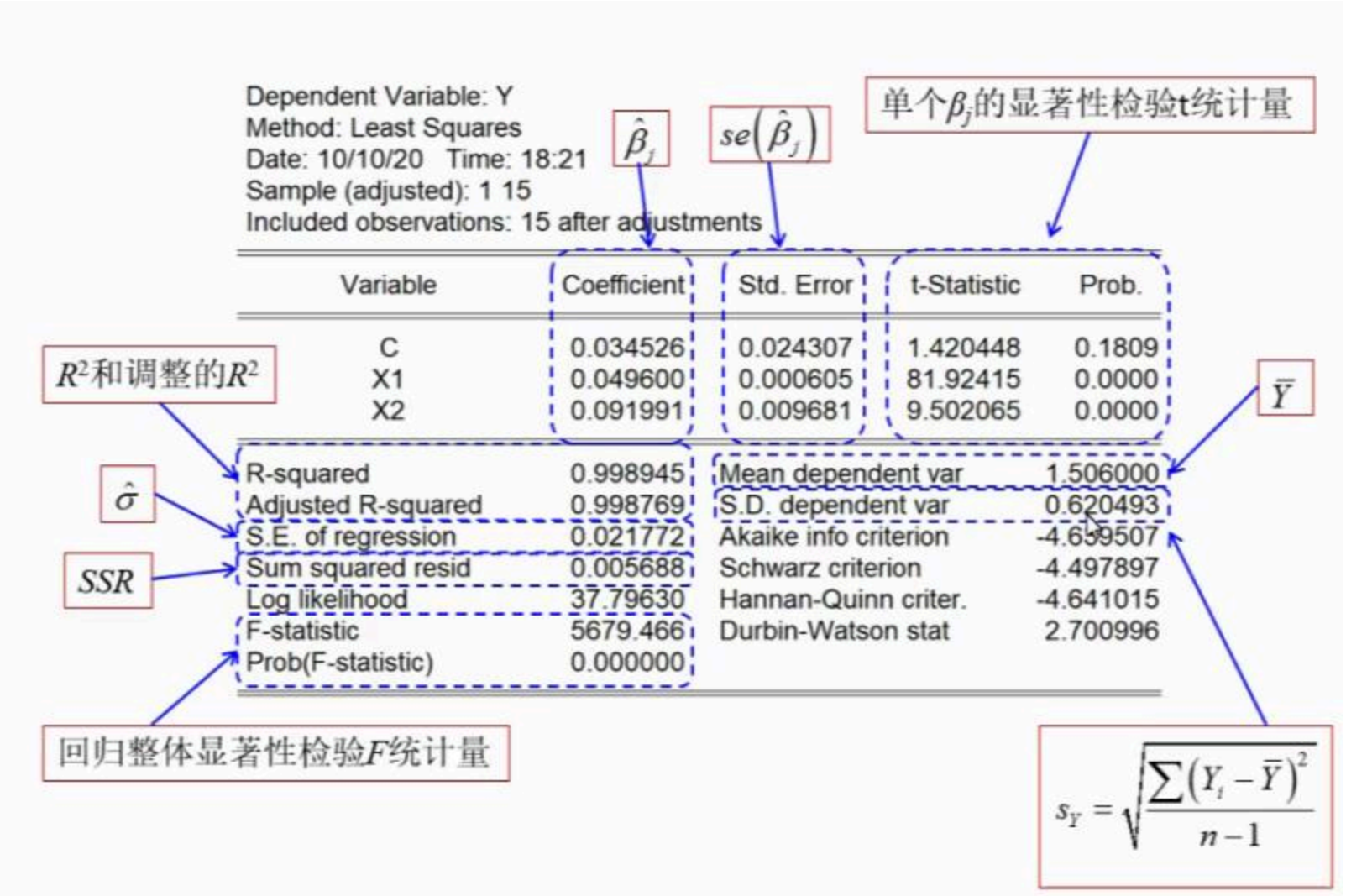
$$F = \frac{SSE/k}{SSR/(n-k-1)} = \frac{R^2/k}{1-R^2/(n-k-1)}$$

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} = \frac{R^2_{ur} - R^2_r/q}{(1-R^2_{ur})/(n-k-1)}$$

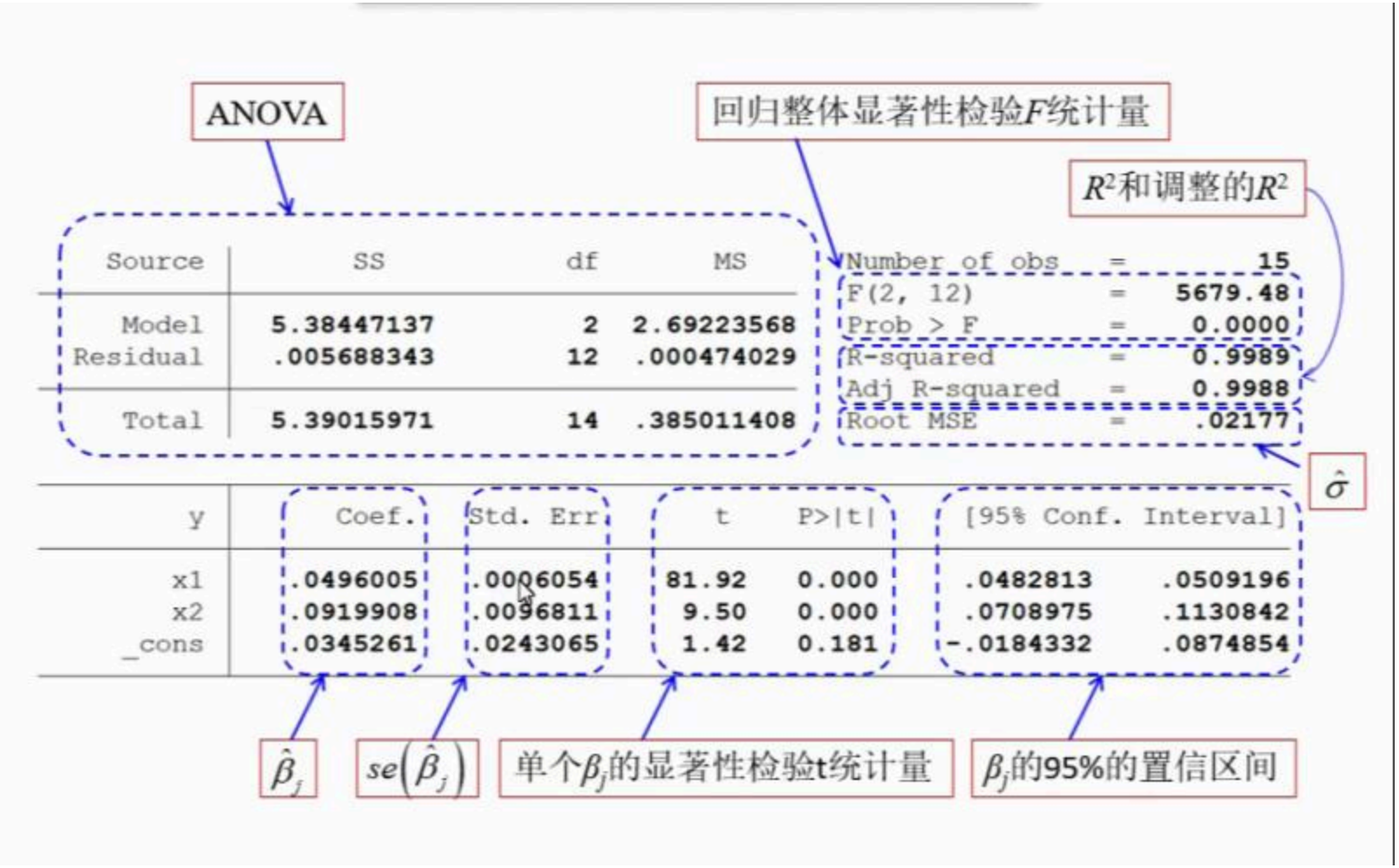
$$\overline{R^2} = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)} = 1 - (1-R^2)\frac{n-1}{n-k-1}$$

$$SER = \widehat{\sigma}^2 = \frac{\sum \widehat{u}_i^2}{n-k-1}$$

EViews结果解读：


Dependent Variable: Y
Method: Least Squares
Date: 10/10/20 Time: 18:21
Sample (adjusted): 1 15
Included observations: 15 after adjustments
$\hat{\beta}_j$
$se(\hat{\beta}_j)$
单个β_j的显著性检验t统计量
Variable Coefficient Std. Error t-Statistic Prob.
C 0.034526 0.024307 1.420448 0.1809
X1 0.049600 0.000605 81.92415 0.0000
X2 0.091991 0.009681 9.502065 0.0000
R^2和调整的R^2
$\hat{\sigma}$
SSR
$\bar{Y}$
R-squared 0.998945
Adjusted R-squared 0.998769
S.E. of regression 0.021772
Sum squared resid 0.005688
Log likelihood 37.79630
F-statistic 5679.466
Prob(F-statistic) 0.000000
Mean dependent var 1.506000
S.D. dependent var 0.620493
Akaike info criterion -4.639507
Schwarz criterion -4.497897
Hannan-Quinn criter. -4.641015
Durbin-Watson stat 2.700996
回归整体显著性检验F统计量
$s_Y = \sqrt{\frac{\sum(Y_i - \bar{Y})^2}{n-1}}$

Stata结果解读：



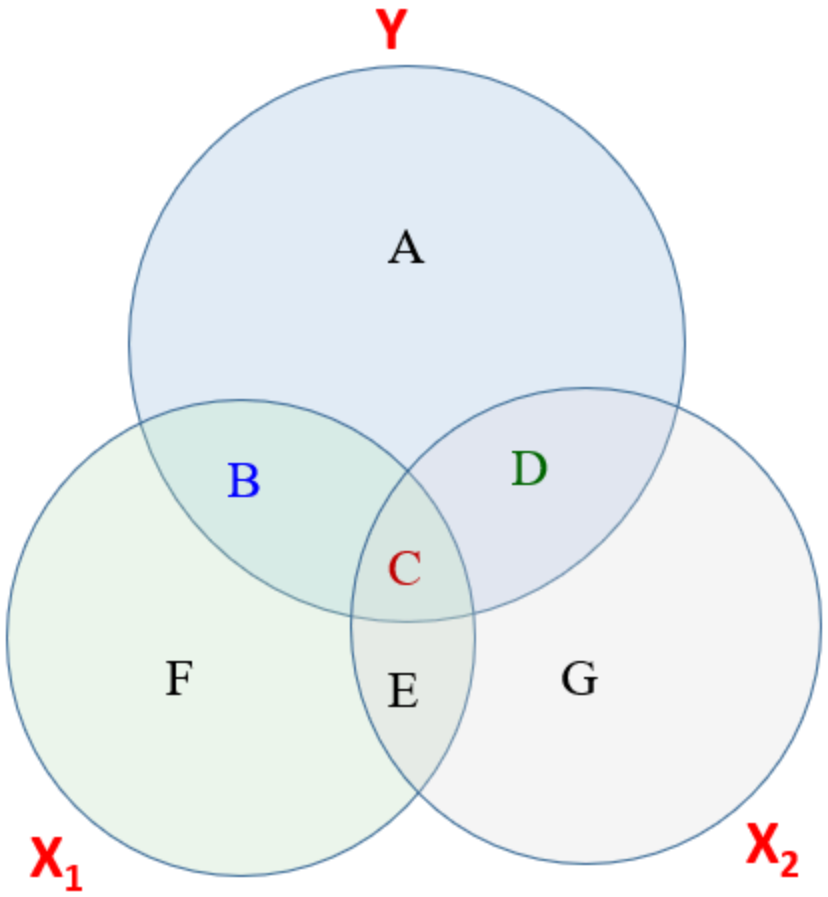
Source	SS	df	MS
Model	5.38447137	2	2.69223568
Residual	.005688343	12	.000474029
Total	5.39015971	14	.385011408

Number of obs	= 15
F(2, 12)	= 5679.48
Prob > F	= 0.0000
R-squared	= 0.9989
Adj R-squared	= 0.9988
Root MSE	= .02177

y	Coef.	Std. Err	t	P>|t|	[95% Conf.	Interval]
x1	.0496005	.0006054	81.92	0.000	.0482813	.0509196
x2	.0919908	.0096811	9.50	0.000	.0708975	.1130842
_cons	.0345261	.0243065	1.42	0.181	-.0184332	.0874854

2. 运用Venn图理解偏回归系数、遗漏变量、多重共线性

(整理自连享会gitee)



令A+B+C+D=1

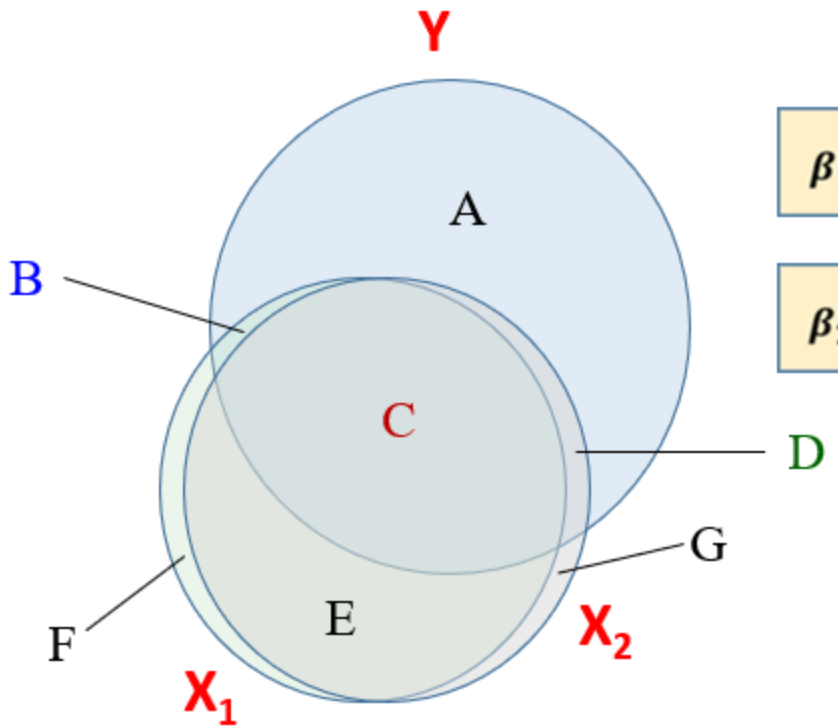
$R^2 = \frac{B+C}{A+B+C+D}$

reg Y X1 → R^2 = B+C

reg Y X2 → R^2 = C+D

reg Y X1 X2 → R^2 = B+C+D

Case



$\beta_1 = \frac{\partial Y}{\partial X1}|_{X2}$

$\beta_2 = \frac{\partial Y}{\partial X2}|_{X1}$

reg Y X1 X2 → R^2 = B+C+D

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

reg Y X1 → $\hat{u}$ = A+D

reg X2 X1 → $\hat{e}$ = D+G

reg $\hat{u}$ $\hat{e}$ → R^2 = D

干净的Y

干净的X2

$D = \widehat{\beta_2}|_{X1}$

$B = \widehat{\beta_1}|_{X2}$

两步法:

第一步，用该解释变量对其他解释变量回归，得到OLS残差；

第二步，用y对第一步的残差回归。

三步法（与两步法等价）：

```
reg Y X1
predict u, res
reg X2 X1
predict e, res
reg u e
```

3. 遗漏变量偏误

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$
$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$$
$$x_2 = \delta_0 + \delta_1 x_1 + v$$
$$y = \beta_0 + \beta_1 x_1 + \beta_2(\delta_0 + \delta_1 x_1 + v) + u = (\beta_0 + \beta_2\delta_0) + (\beta_1 + \beta_2\delta_1)x_1 + (\beta_2 v + u)$$

偏误方向即$\beta_2\delta_1$的正负：

	$Corr(x1, x2) > 0$	$Corr(x1, x2) < 0$
$\beta_2 > 0$	$+$	$-$
$\beta_2 < 0$	$-$	$+$

Wooldridge 6e, chap3:

(1)Example 3.4 Determinants of College GPA

```
corr colGPA hsGPA ACT
reg colGPA hsGPA ACT
reg colGPA ACT
```

(2)Example 3.6 Hourly Wage Equation

(3)Problem8

4. 多重共线性的识别与处理

当存在分组时，如果放入全部组别，Stata会自动删除一个组，以避免完全共线性问题。

识别方法：相关系数矩阵、方差膨胀因子（VIF） `estat vif`

处理方法：删除或重新定义变量、逐个放入

5. （不要求掌握）高级估计方法：岭回归(Ridge Regression)

- 大数据表现为“高维数据”，即特征向量的维度远大于样本容量。
- 在传统实证研究中，样本量一般远大于变量个数：在上市公司的研究中，上市公司的数量大于回归中使用的特征变量个数——使用OLS没有问题
- 但如果是某研究收集了100个病人的信息，其中每个病人均有2万条基因（即2万个特征变量），需要研究哪些基因导致了某种疾病。在这种高维数据的情况下，如果沿用OLS回归，就非常容易出现变量间的**严重多重共线性问题**。

- OLS Regression:

$$f = \sum_{i=1}^{n} (y_i - X\hat{\beta})^2$$

- Ridge Regression:

$$f = \sum_{i=1}^{n} (y_i - X\hat{\beta})^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

- Lasso Regression:

$$f = \sum_{i=1}^{n} (y_i - X\hat{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

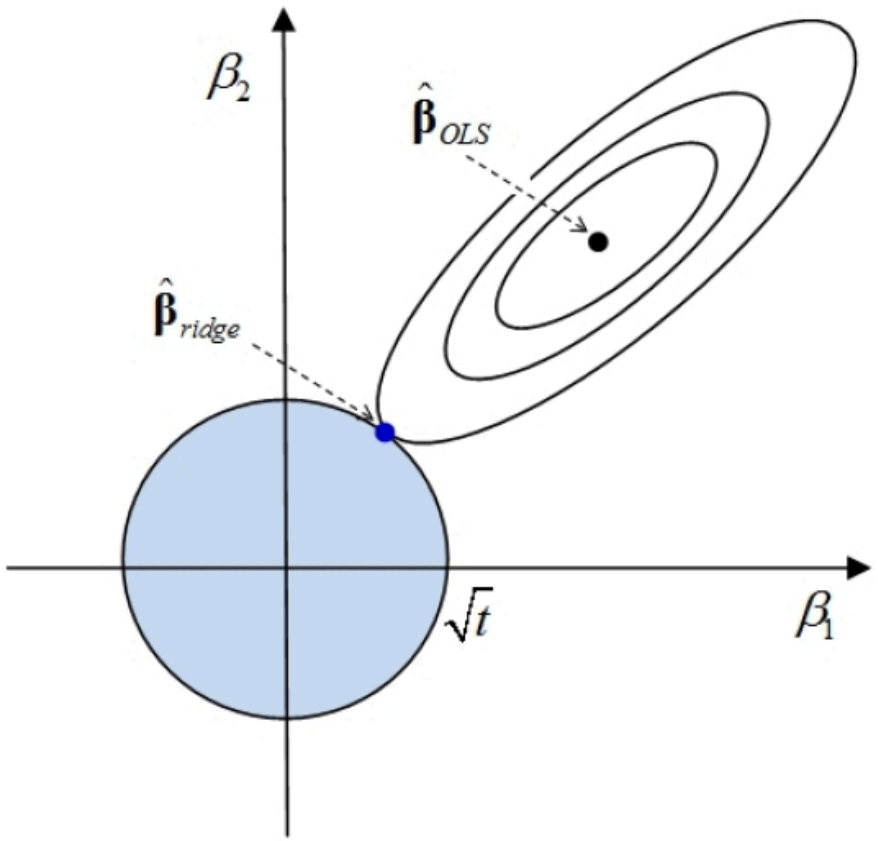


图 9.2　岭回归的约束条件示意图

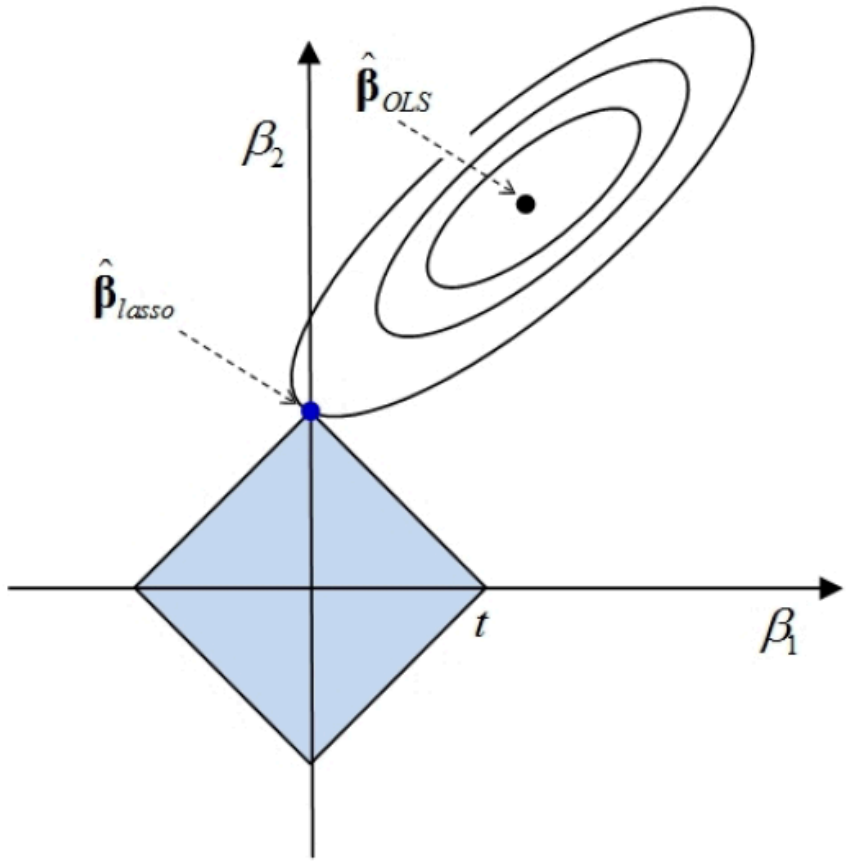


图 9.3　Lasso 的约束条件示意图

Stata command:

- Ridge Regression: `ridgereg`, `rxridge`
- Lasso Regression: `lassopack` (`lasso2`, `cvlasso`, `rlasso`)
- Elasti Net: `elasticregress`

参考资料：

- 陈强编著.机器学习及Python应用.高等教育出版社.2021 **chap9**
- 郭峰等编著.机器学习与社会科学应用.上海财经大学出版社.2024 **chap2**
- Stata：拉索回归和岭回归-(Ridge,-Lasso)-简介
- Stata：拉索开心读懂-Lasso入门
- 图解Lasso系列A：Lasso的变量筛选能力