计 量 经 济 学

Econometrics

第四讲 模型设定

(教材第6、7章)

回顾：OLS的基本假设

假设1：回归模型是线性的，模型设定无误且含有误差项

假设2：误差项总体均值为零 $E(\varepsilon_i)=0$

假设3：所有解释变量与误差项都不相关 $Cov(X_i, \varepsilon_i)=0$

假设4：误差项观测值互不相关（无序列相关性） $Cov(\varepsilon_i, \varepsilon_j)=0$

假设5：误差项具有同方差(不存在异方差性) $Var(\varepsilon_i)=\sigma^2$

假设6：任何一个解释变量都不是其他解释变量的完全线性函数（不存在完全多重共线性）

基本假设1：模型设定无误

- 什么是正确的方程?
  - 正确的解释变量
  - 正确的函数形式
  - 正确的随机误差形式

主要内容

- 模型设定一：选择正确的解释变量
- 模型设定二：选择正确的函数形式

怎样选择解释变量?

- 最重要的选择依据：(经济)理论判断
  - 某个变量应该作为解释变量，即便统计上是不显著的
  - 例：若研究某商品的需求量，应选择那些解释变量?
- 若理论上不明确，则可采用统计方法来判断
  - 遗漏变量 (omitted variable)
  - 不相干变量 (irrelevant variable)

微观经济理论

遗漏变量的后果

$$Y = X_1\beta_1 + \qquad \varepsilon^*$$

如何证明？

- 设定偏误：参数估计量有偏且非一致，方差变小

遗漏变量的后果

$$Y = X_1\beta_1 + X_2\beta_2 + \varepsilon = X_1\beta_1 + \varepsilon^*$$

假设3：所有解释变量与误差项都不相关 $\mathbf{Cov(X_i, \varepsilon_i)=0}$

遗漏变量的后果

$$Y = X_1\beta_1 + \qquad \varepsilon^*$$

- 设定偏误：参数估计量有偏且非一致，方差变小

- 如果方程遗漏一个相关变量，则会出现：

（1）无法从方程中获取遗漏变量的参数估计值。

（2）方程中其余变量的参数估计也会出现偏误。

请仔细阅读教材p95：遗漏变量偏误示例

例：估计鸡肉需求方程(Table 6-2)

$$Y_t = \beta_0 + \beta_1 PC_t + \beta_2 YD_t + \varepsilon_t$$

Sample: 1974 1992
Included observations: 19

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	40.00663	3.762011	10.63437	0.0000
PC	-0.363653	0.091215	-3.986772	0.0011
YD	0.303357	0.013405	22.63077	0.0000

R-squared	0.988404	Mean dependent var	54.76158
Adjusted R-squared	0.986955	S.D. dependent var	11.43216

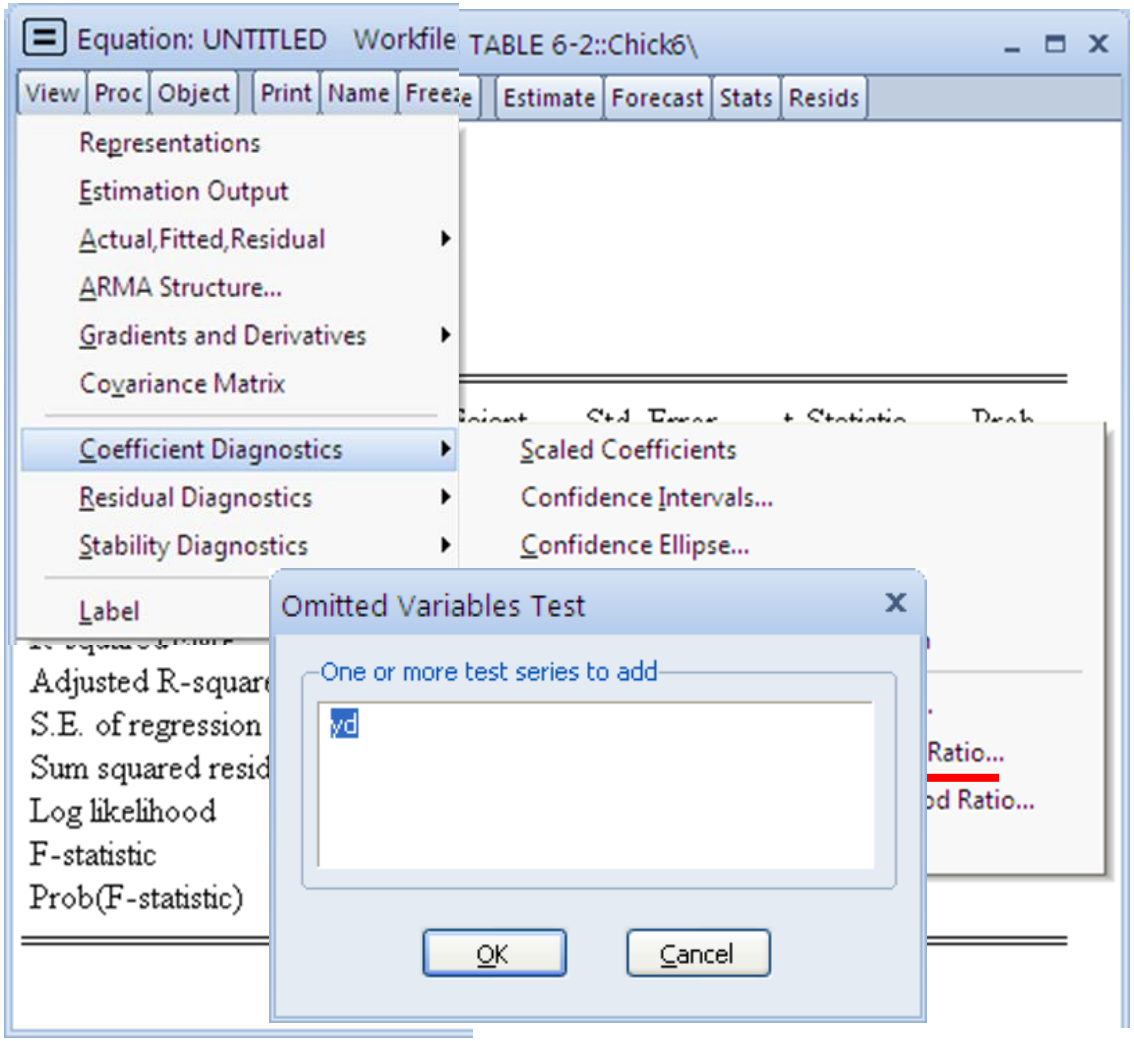
其中，PC是价格，YD是人均可支配收入。假设遗漏了变量*YD（人均可支配收入）*则回归结果为(存在设定误差):

Sample: 1974 1992
Included observations: 19

不符合预期

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-21.13566	14.59173	-1.448469	0.1657
PC	1.394630	0.266365	5.235777	0.0001

R-squared	0.617232	Mean dependent var	54.76158
Adjusted R-squared	0.594717	S.D. dependent var	11.43216



原假设：YD不是遗漏变量。

结论：拒绝原假设，接受YD是遗漏变量。

Omitted Variables Test
Equation: UNTITLED
Specification: Y C PC
Omitted Variables: YD

	Value	df	Probability
t-statistic	22.63077	16	0.0000
F-statistic	512.1516	(1, 16)	0.0000
Likelihood ratio	66.43910	1	0.0000

加入不相干变量的后果

$$Y = X_1\beta_1 + \qquad \varepsilon$$

- 参数估计量无偏但非有效(方差增大)，t检验失效

加入不相干变量的后果

$$Y = X_1\beta_1 + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon^{**}$$

加入不相干变量的后果

$$Y = X_1\beta_1 + \qquad \varepsilon$$

- 参数估计量无偏但非有效(方差增大)，*t*检验失效
- 当方程中存在不相干变量时，通常调整判定系数adj-R^2会减小。

请仔细阅读教材p100：误选不相干变量的实例

例：估计鸡肉需求方程(Table 6-2)

$$Y_t = \beta_0 + \beta_1 PC_t + \beta_2 YD_t + \varepsilon_t$$

Sample: 1974 1992
Included observations: 19

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	40.00663	3.762011	10.63437	0.0000
PC	-0.363653	0.091215	-3.986772	0.0011
YD	0.303357	0.013405	22.63077	0.0000

R-squared	0.988404	Mean dependent var	54.76158
Adjusted R-squared	0.986955	S.D. dependent var	11.43216

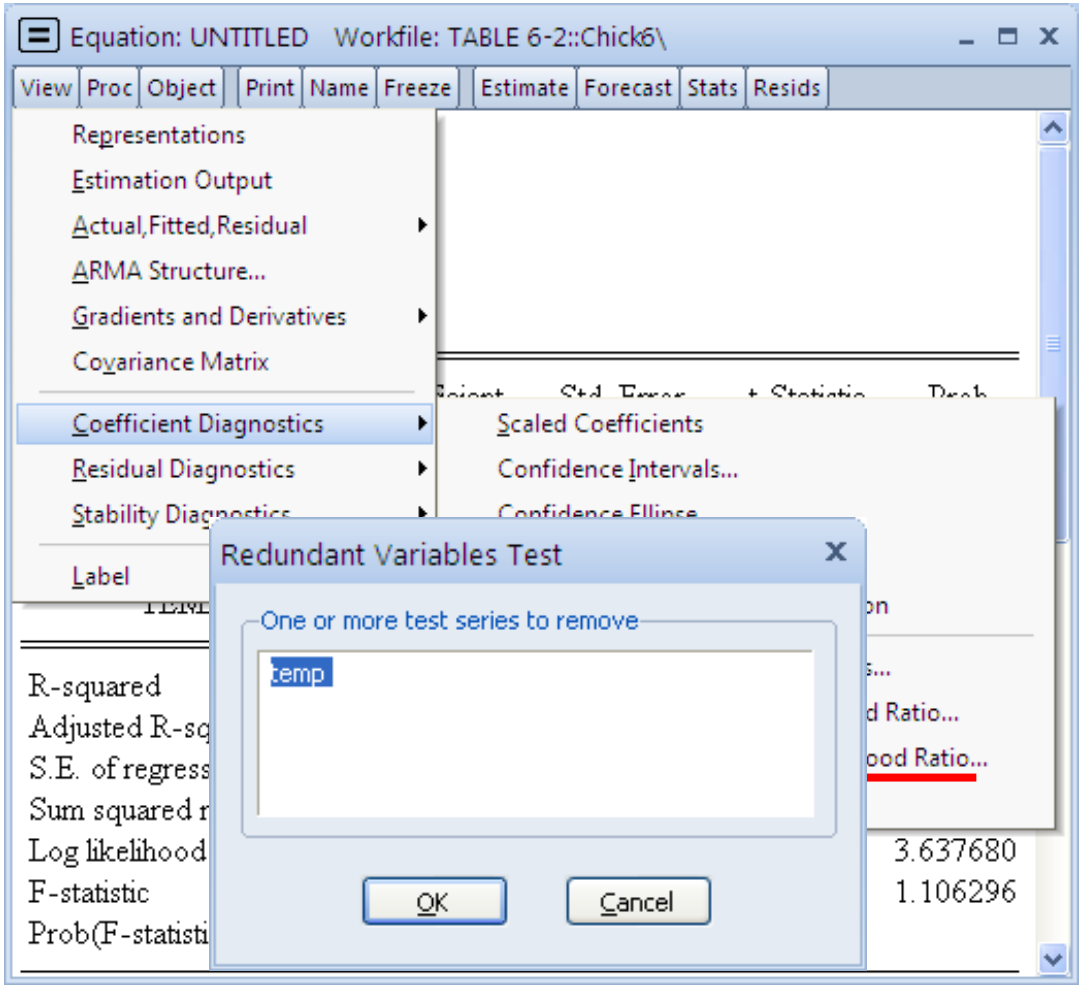
假设误选了不相干变量*TEMP*（气温）则回归结果为：

Sample: 1974 1992
Included observations: 19

Variable	Coefficient	Std		
C	40.39477	3.93	10.27714	0.0000
PC	-0.368092	0. 849	-3.922184	0.0014
YD	0.300411	0.014938	20.11107	0.0000
TEMP	0.009469	0.018915	0.500647	0.6239

不显著

R-squared	0.988595	Mean dependent var	54.76158
Adjusted R-squared	0.986314	S.D. dependent var	11.43216



Redundant Variables Test
Equation: UNTITTLED
Specification: Y C PC YD TEMP
Redundant Variables: TEMP

	Value	df	Probability
t-statistic	0.500647	15	0.6239
F-statistic	0.250647	(1, 15)	0.6239
Likelihood ratio	0.314863	1	0.5747

原假设：TEMP是冗余变量。

结论：接受原假设，TEMP是冗余变量。

模型选择准则

- 判定系数

$$R^2 = 1 - \frac{RSS}{TSS}$$

- 调整的判定系数

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

K为待估参数的个数

原则上，增加的变量非冗余变量时，adj-R^2应该增大。

模型选择准则（两个或更多模型的选择）

- 赤池信息准则(AIC)和施瓦茨信息准则(SC)
- k为待估参数的个数。

$$AIC = \frac{2k}{n} + \ln\left(\frac{RSS}{n}\right)$$

$$SC = \frac{k}{n}\ln n + \ln\left(\frac{RSS}{n}\right)$$

AIC和SC准则对增加解释变量加大了惩罚，其中SC的惩罚比AIC更严厉。

AIC和SC的判定准则：相对而言，AIC和SC的值越低的模型越好。

Sample: 1974 1992
Included observations: 19

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-21.13566	14.59173	-1.448469	0.1657
PC	1.394630	0.266365	5.235777	0.0001

R-squared	0.617232	Mean dependent var	54.76158
Adjusted R-squared	0.594717	S.D. dependent var	11.43216
S.E. of regression	7.277929	Akaike info criterion	6.906870
Sum squared resid	900.4602	Schwarz criterion	7.006285

这个模型更好

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	[illegible]	3.762011	10.63437	0.0000
PC	-0.363653	0.091215	-3.986772	0.0011
YD	[illegible]303357	0.013405	22.63077	0.0000

R-squared	0.98[illegible]4	Mean dependent var	54.76158
Adjusted R-squared	0.986955	S.D. dependent var	11.43216
S.E. of regression	1.305729	Akaike info criterion	3.515339
Sum squared resid	27.27884	Schwarz criterion	3.664461

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	40.39477	3.930548	10.27714	0.0000
PC	-0.368092	0.093849	-3.922184	0.0014
YD	0.300411	0.014938	20.11107	0.0000
TEMP	0.009469	0.018915	0.500647	0.6239

R-squared	0.988595	Mean dependent var	54.76158
Adjusted R-squared	0.986314	S.D. dependent var	11.43216
S.E. of regression	1.337423	Akaike info criterion	3.604030
Sum squared resid	26.83051	Schwarz criterion	3.802860

四个重要的模型设定准则

- 理论：变量在方程
- t 检验：解释变量参数的估计值在预期假设下是不是显著的？
- 调整判定系数adj-R^2或AIC和SC：将变量加入方程后，整体拟合优度是否有所改善？
- 偏误：将变量加入方程后，其他变量参数是否有显著变化？

不清的，从理论上看是不

不能简单地将一个t值不显著的变量从方程中排除

误用模型设定准则的实例

巴西咖啡的价格

茶叶价格

$$\hat{C}OFFEE = 9.1 + 7.8P_{bc} + 2.4P_t + 0.0035Y_d$$

标准误差 (15.6) (1.2) (0.001)

$t =$ 0.5 2.0 3.5

$\bar{R}^2 = 0.6 \quad N = 25$

美国对巴西咖啡的需求量

美国可支配收入

$$\hat{C}OFFEE = 9.3 + 2.6P_t + 0.0036Y_d$$

(1.0) (0.0009)

$t =$ 2.6 4.0

$\bar{R}^2 = 0.61 \quad N = 25$

误用模型设定准则的实例

哥伦比亚咖啡的价格

巴西咖啡的价格

$$\hat{C}OFFEE = 10 + 8.0P_{cc} - 5.6P_{bc} + 2.6P_t + 0.003Y_d$$

标准误差 (4.0) (2.0) (1.2) (0.001)

$t =$ 2.0 -2.8 2.0 3.0

$\bar{R}^2 = 0.65 \quad N = 25$

请仔细阅读教材p106：选择解释变量的实例

模型设定搜索

- 数据挖掘 (data mining)
  - 适当的数据挖掘也许有助于揭示经验规律
  - 不适当的数据挖掘，比什么都不据严刑拷打，它就会屈打成招

- 敏感性分析：稳健性(robust)分析
  - 几乎所有学术论文的必备步骤和分析内容
  - 稳健的含义：某种结果对于各种模型设定、变量定义、数据子集都是显著的(或不显著的)

怎样选择函数形式?

- 第7章 模型设定：函数形式的选择(随机抽点)
  - 7.1节：常数项的应用和解释
  - 7.2节：备选函数形式
  - 7.5节：选择错误函数形式存在的问题

本讲小结

- 遗漏变量的危害是什么?
- 不相干变量的后果是什么?
- 模型设定的四个重要准则是什么?
- 回归方程能省略常数项吗?
- 如何选择回归模型的函数形式?
- 在回归方程中，是否需要剔除所有不显著的解释变量?

作业

- P108：习题2、6
- 第七章（P128）：习题1、3、4、5

怎样选择函数形式?

- 错误函数形式的后果
  - 影响解释变量的显著性
  - 解释变量可能有非预期的符号
  - 严重影响模型解释和变量预测

函数形式的选择

- 不含常数项的回归
- 回归模型的函数形式
- 函数形式的选择

不含常数项的回归

- 不含常数项的回归：

$$Y_i = \beta_1 X_i + \varepsilon_i$$

不要信赖 (分析) 常数项的估计值

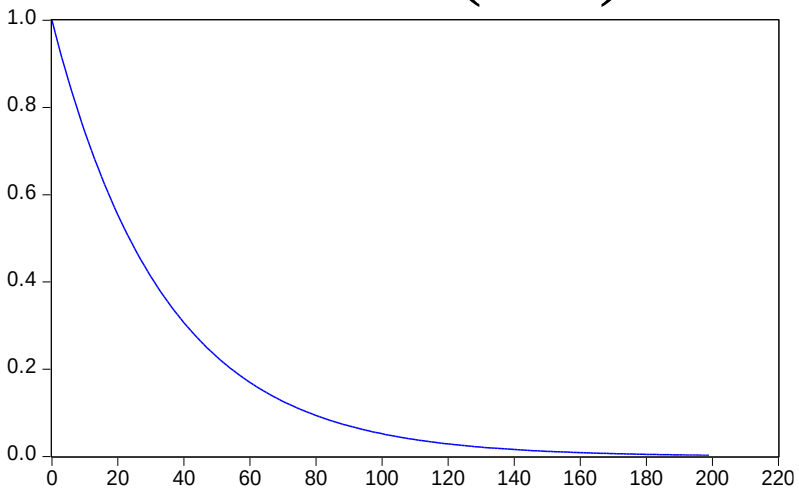
- 可以证明：
  - 残差均值不一定为0
  - 拟合优度的判定系数可能出现负值

除非有非常强的先验预期，否则还是采取含有常数项的模型为好；即使先验预期为无常数项模型，仍可使用含常数项的模型，再检验其常数项在统计上等于0即可。

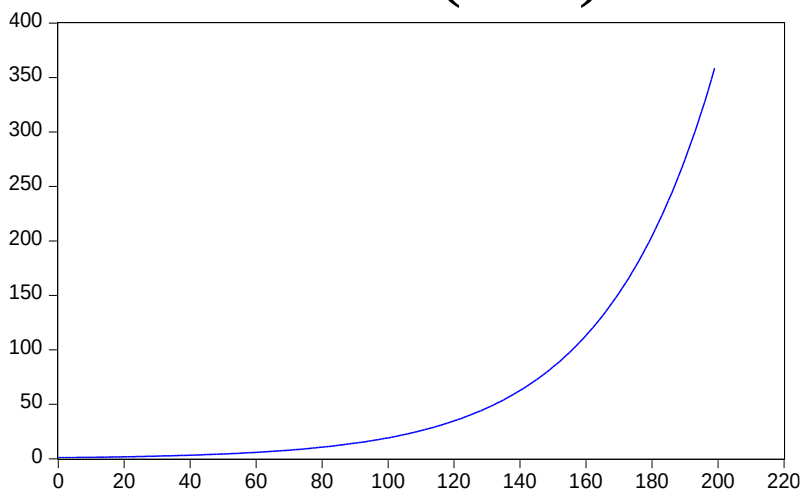
备选函数形式

- 线性回归模型能否满足所有问题的需求?
- 例1：债券期限对债券价格的影响(现值)、存款期限对存款本息的影响(终值)

$$Y = FV(1+r)^{-t}$$

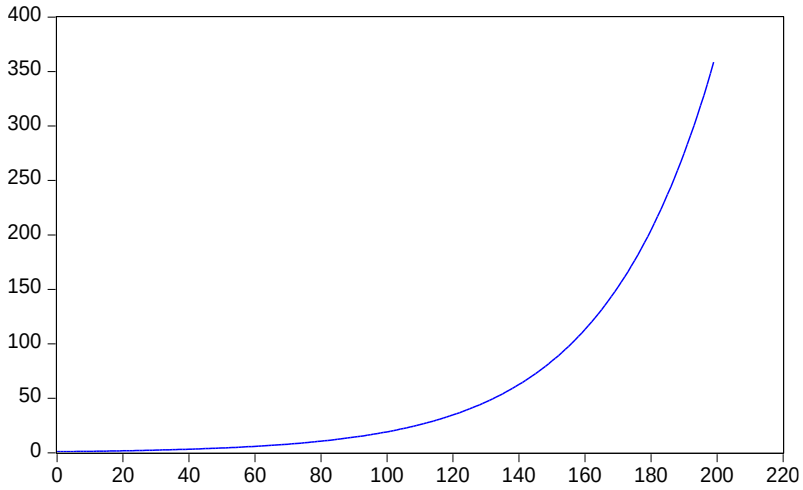


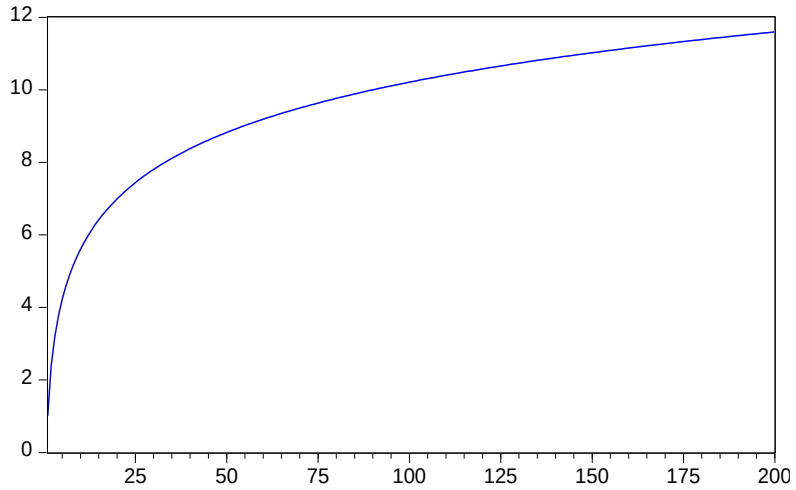
$$Y = PV(1+r)^{t}$$



备选函数形式

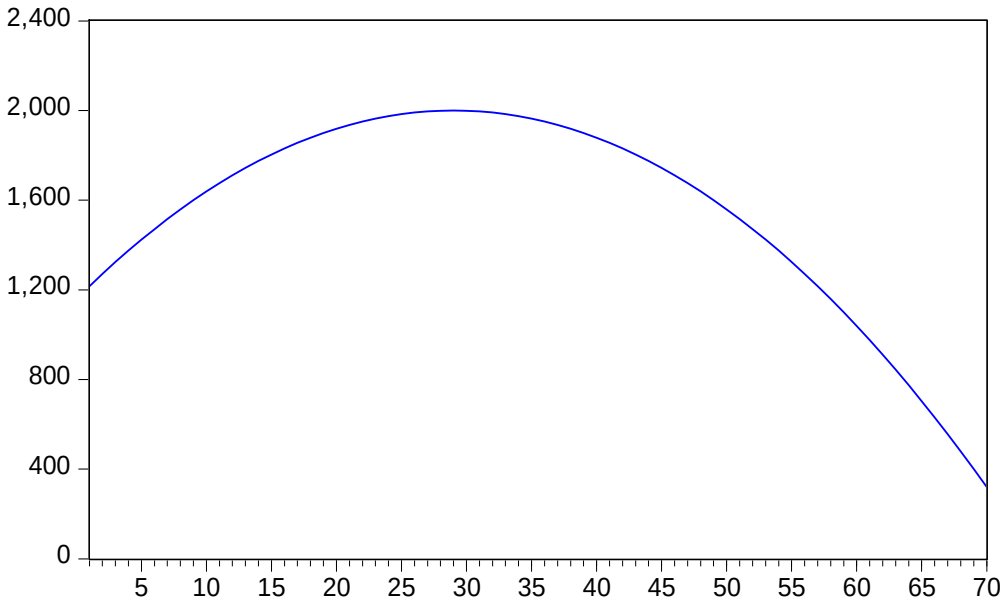
❖ 例2：投入对总产量的影响(边际产量递增、递减)





备选函数形式

❖ 例3：工人年龄对其收入的影响



备选函数形式

- 可线性化的非线性函数形式
  - 指数函数
  - 对数函数
  - 反函数形式
  - 多项式形式

回归模型的函数形式

- 指数函数

$$y_t = ae^{bx_t+\varepsilon_t}$$

对上式等号两侧同取自然对数，得

$$\ln y_t = \ln a + b\,x_t + \varepsilon_t$$

令$\ln y_t = y_t^*$, $\ln a = a^*$, 则

$$y_t^* = a^* + bx_t + \varepsilon_t$$

变量y_t^* 和x_t已变换成为线性关系。

回归模型的函数形式

❖ 半对数线性模型

考虑如下复利公式：

偏回归系数表示增长率，即给定X的绝对变化引起的Y的相对变化

$$Y_t = Y_0(1+r)^t e^{\varepsilon_t}$$

可转化成半对数模型

$$\ln Y_t = \ln Y_0 + t\ln(1+r) + \varepsilon_t$$

$$= \beta_0 + \beta_1 t + \varepsilon_t$$

时间变量t称为趋势变量

回归模型的函数形式

- 柯布-道格拉斯(Cobb-Douglas)生产函数：

$$Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} e^{\varepsilon}$$

两边取对数有：$\ln Y = \ln \beta_0 + \beta_1 \ln X_1 + \beta_2 \ln X_2 + \varepsilon$

偏回归系数表示弹性，即给定X的百分比变化引起的Y的百分比变化。

回归模型的函数形式

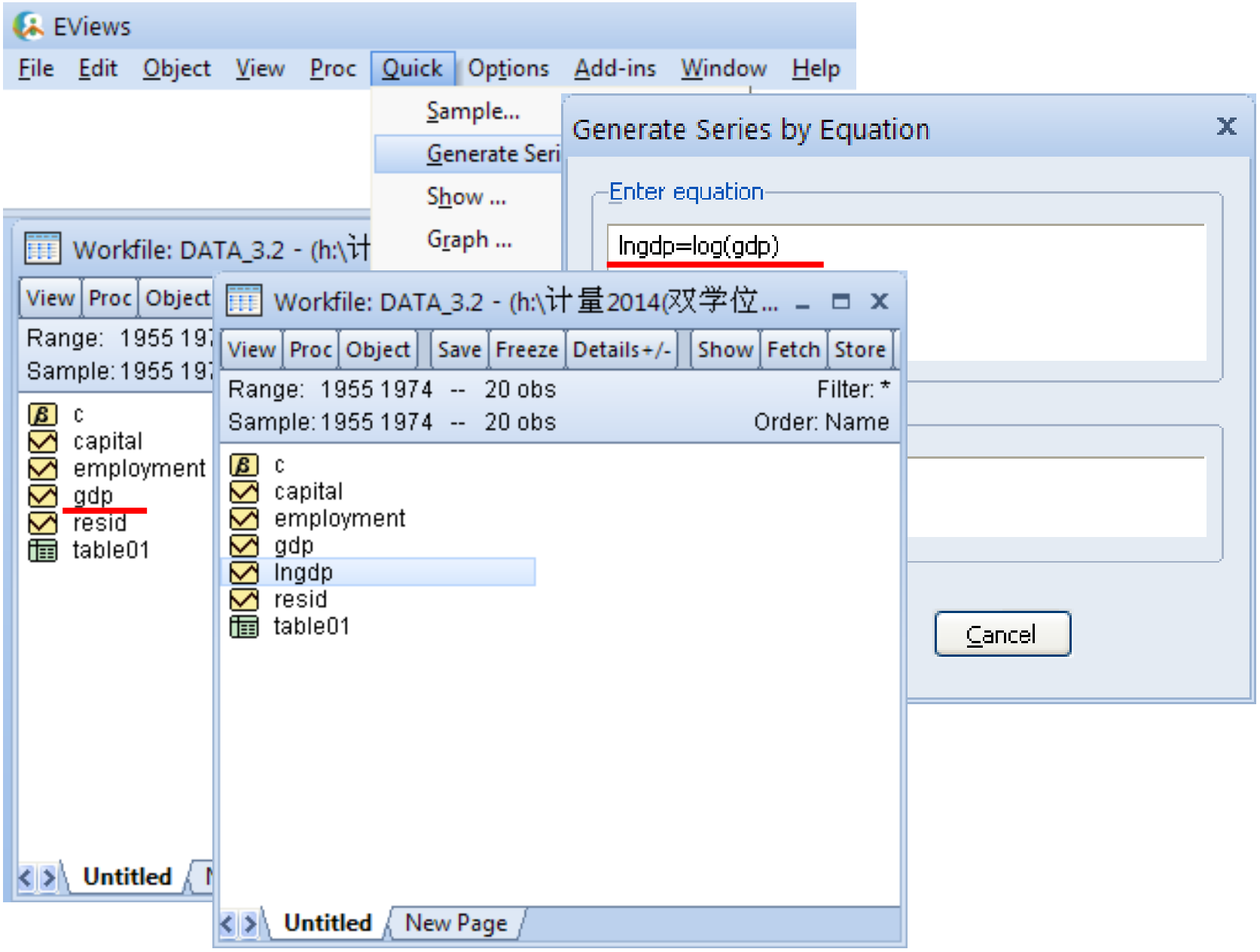
- 反函数形式

$$Y_i = \beta_0 + \beta_1 \frac{1}{X_i} + \varepsilon_i$$

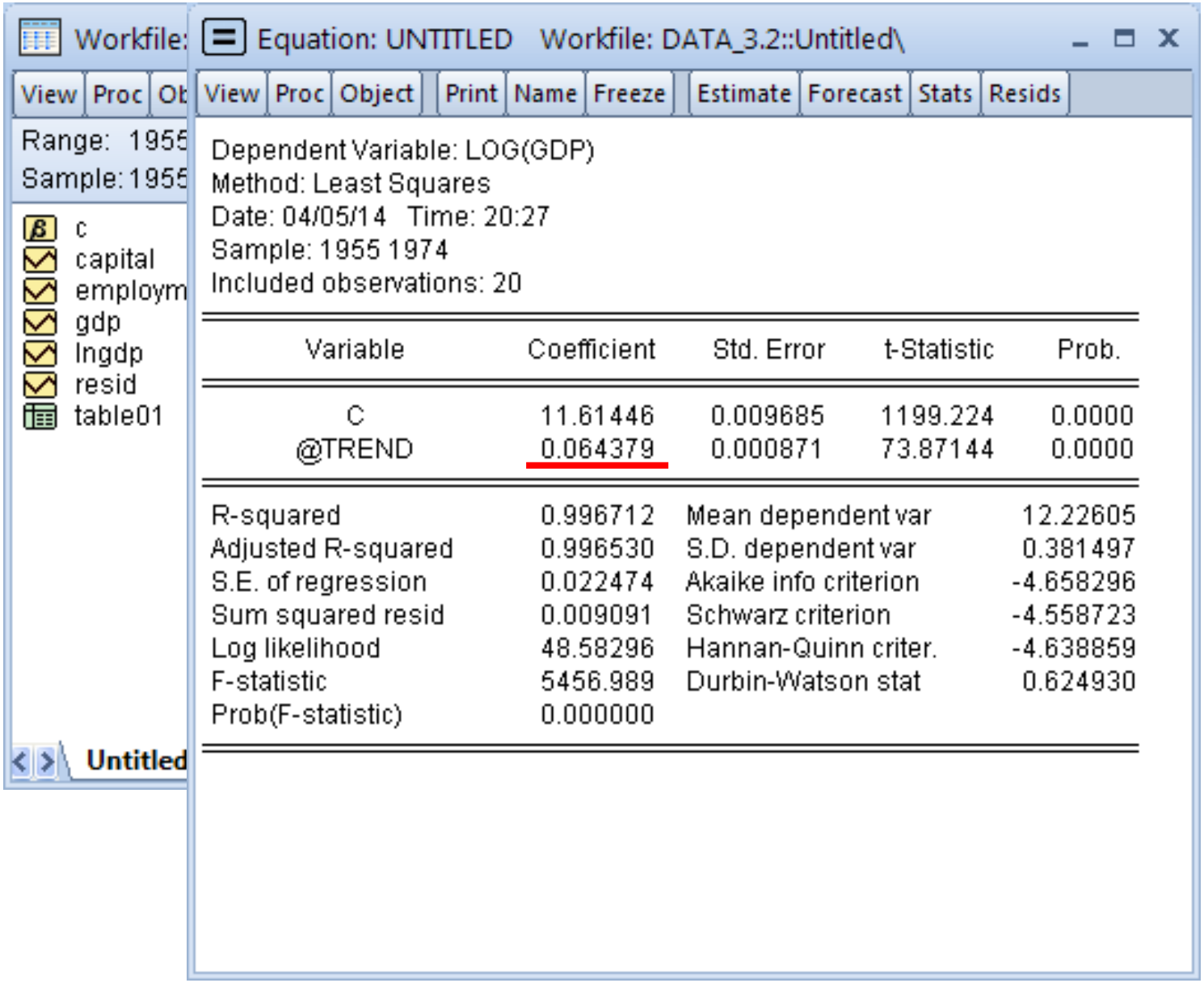
- 多项式形式

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$$

EViews演示：变量转换或生成新的变量



EViews演示：GDP增长率



回归模型的函数形式

- 如何选择函数形式
  - 选择经济理论给出的特定函数形式
  - 所选模型的参数应满足一定的先验预期
  - 当多个模型能很好地拟合数据时，研究者往往选择调整的判定系数较高或者AIC和SC较低的模型。然而，当被解释变量*Y*被变换时，这些指标不能比较

作业

- 第七章（P128）： 习题1、3、5、8